PHONEME ADJUSTMENT

IN ENHANCED SPEECH

THESIS

Nadeem A. Bashir, Flt.Lt. PAF

AFIT/GE/ENG/89M-2

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

AFIT/GE/ENG/89M-2

PHONEME ADJUSTMENT

IN ENHANCED SPEECH

THESIS

Nadeem A. Bashir, Flt.Lt. PAF

AFIT/GE/ENG/89M-2

AFIT/GE/ENG/89M-2

PHONEME ADJUSTMENT

IN ENHANCED SPEECH

THESIS

Presented to the Faculty of School of Engineering

of the Air Force Institute of Technology

Air University

In Partial Fulfillment of the

Requirements for the Degree of

Master of Science in Electrical Engineering

Nadeem A. Bashir, Flt.Lt. PAF

March 1989

Acknowledgments

This work is dedicated to all the loved ones, specially my mother whose prayers and moral support made this work possible.

Also, special thanks to my thesis advisor, Dr.Mathew Kabrisky, for his inspiration, knowledge and freedom of work he bestowed upon me during the thesis work

Table of Contents

List of Figures

AFIT/GE/ENG/89M-2

Abstract

A system was developed to enhance the quality and intelligibility of speech which had been pre-processed by a Speech Enhancement Unit (SEU) at RADC Griffis AFB. The system processes the speech in the frequency domain. A Hamming window with 50% overlap was applied to the time waveform and a 512-point Discrete Fourier Transform (DFT) was computed. The amplitude spectrum of voiced regions was smoothed in order to reduce the effects of noise. Frequencies above 2.5 KHz were enhanced as they had been attenuated by SEU. Harmonics of the glottal pitch frequency of voiced speech and peaks of unvoiced speech were selected to further reduce the noise effects. The harmonics selected were not necessarily the exact harmonics of the glottal frequency. The two neighboring frequency points were checked and the maximum of those three points was selected instead of the exact glottal harmonic. Speech was reconstructed using amplitude phase, and frequency of the harmonics/peaks selected. The reconstructed speech had much better quality and improved SNR. SPIRE (Speech and phonetics Interactive Research Environment) and ILS (Interactive Laboratory System) software packages were used for visual analysis of the amplitude spectrum. The system was implemented in FORTRAN 77 on a VAX 11/780 machine. Keywords: Speech processing, [illegible]. (KR)

PHONEME ADJUSTMENT

IN ENHANCED SPEECH

I. Introduction

The redundancy inherent in speech makes possible the ability of human listeners to detect and understand speech even when it is severely distorted or heavily obscured by noise. However, the human listener cannot listen to speech under degraded conditions for long periods of time without suffering auditory fatigue. This reduces the ability of listener to recognize speech and understand it (2:1-1). In order to reduce the auditory fatigue and to increase the intelligibility of the noisy speech, enhancement of speech is often employed.

The main objective of the speech enhancement is to ultimately improve one or more perceptual aspects of speech, such as overall quality, intelligibility, or degree of listener fatigue. The major motivating force for speech enhancement in military is to correct speech jammed

by enemy signals and to enhance intercepted enemy speech which is often highly degraded by noise.

Background

The problem of enhancing speech degraded by noise has received considerable attention in recent years (1). The objective of speech enhancement may be human listening or input to a speech/speaker recognition system. In the case where the objective is human listening, the perceptual aspects of the enhanced speech, as quality, intelligibility, and pleasantness become important. Most of the investigators have considered these aspects of speech enhancement (4;5).

Apart from direct human listening, speech enhancement also has important applications in the area of speech/speaker recognition by machines. Often the basis of this system is a parametric model, the parameters of which are extracted from the input speech. When speech is degraded due to additive noise, the estimated parameters are distorted, resulting in the deterioration of recognition performance.

One of the approaches of speech enhancement is the resynthesis of speech from the peaks of spectral

magnitude. The reconstruction of speech is done through sinusoidal waveforms because speech can be modelled as a sum of sine waves. The number of peaks required to maintain the quality increases as the glottal pitch frequency of the speech decreases. As few as 16 peaks are required for high-pitched female speech, while as many as 40 peaks are required for low-pitched male speech (5:27.6.2). The major difficulty in synthesizing speech from this method is the time-variablity of the number and location of peaks estimated in each frame. Hence frame-to-frame peak matching becomes the critical part in accurate reconstruction of the speech (5:27.6.3). Accurate frequency estimation is also necessary for high quality reconstruction of speech using this method.

Another approach to speech enhancement is to use the fact that waveforms of voiced sounds are approximately periodic. The periodicity of a time waveform translates itself in the frequency domain as harmonics of fundamental frequency corresponding to the period of time waveform (6:4). Since the energy of a periodic signal is concentrated in bands of frequencies and the interfering signals in general have energy over the entire frequency band, the use of adaptive comb filtering can enhance the speech considerably. Separation of speech from interfering noise by means of harmonic selection can be regarded as a

frequency-domain implementation of the adaptive comb filter (7:911). This approach, however does not apply to the unvoiced speech because frequency content of unvoiced speech is not harmonically related to each other.

Problem

Rome Air Development Center (RADC) at Griffis AFB is carrying out research and development of speech enhancement techniques that would be helpful to both the human listener and speech recognition devices. RADC has developed a device for this purpose called Speech Enhancement Unit (SEU). The SEU is used to enhance speech by eliminating three kinds of noise: broadband, impulse, and stationary or sweeping tones. This enhancement process, however affects different phonemes differently. Hissing sounds called fricatives (e.g. [s] in Surface), suffer more than the other phonemes. This results in degradation of both quality and intelligibility of speech. Adjustment of phonemes in the SEU-enhanced speech is required so that this problem is rectified and the intelligibility and quality of the speech is increased.

Scope

The processing of the speech is carried out in frequency domain. For this purpose 512-point Discrete Fourier Transform (DFT) is taken and then all the processing is carried out on that DFT. After smoothing the DFT, proper harmonics of the glottal pitch frequency for the voiced speech and peaks of unvoiced speech are taken and speech is reconstructed using amplitude, phase, and frequency of the selected harmonics/peaks.

Approach

The approach is outlined as follows. First the Hamming window with 50% overlap is applied to the time waveform. Next 512-point DFT is computed and the spectrum of all the frames is smoothed (explained later). Then high frequencies are amplified because the SEU processed speech had been processed by a low pass filter with a cutoff at 2.5 KHz. Next, harmonics of glottal frequency for voiced speech and peaks of unvoiced speech are selected. Harmonics selected are not necessarily the exact harmonics of the glottal frequency. The neighboring frequency points are checked and if any of those points has higher amplitude than the exact glottal harmonic then that frequency is selected instead of the harmonic. Speech is synthesized using amplitude, phase, and frequency of the harmonics/peaks selected. Finally the synthesized

speech is averaged to reduce to the original number of frames (doubled because of Hamming overlays) and the output waveform is normalized to make it into integer*2 data type. Integer*2 data type is an integer that can have values in the range -32,768 through 32,767. An integer*2 value takes two bytes of storage.

SPIRE (Speech and Phonetics Interactive Research Environment) (6:6-12) and ILS (Interactive Laboratory System) (8:1) software packages were used to analyze different stages of processing of the spectrum and to develop rules for processing the speech.

Sequence of Presentation

Chapter two gives the hardware and software environments used to digitize and process the speech. Chapter three presents the speech processing system. Details are given for each module of the algorithm developed for processing the speech. Chapter four presents the results of the speech processing carried out on different speech files. Chapter five provides conclusions and recommendations for further application of the processing system. The Appendices contain additional results and the computer program.

II. Development Environment

Introduction

The purpose of this chapter is to introduce the software and hardware components used to develop the processing system. The chapter is divided into four sections. The first section describes the speech digitizing system. The next section tells about the programming language and computer system used for it. The third section describes the displays used from SPIRE, an advanced speech analysis program. The last section describes ILS, a software package, which was also used to obtain displays for visual analysis of speech files.

Speech Digitizing

All speech processing systems require an analog to digital (A/D) and a digital to analog (D/A) converter to sample and digitize the speech. This was provided by audio data conversion system (of the Digital Sound Corporation) DSC-200. The DSC-200 has a maximum sampling rate of 50 KHz and maximum conversion rate of 1,600,000 bytes per second (3). The DSC-200 digitizes speech as frames of 256 point each in integer*2 format which can have values in the range -32,768 through 32,767 . All the speech files were sampled at 16 KHz. This 16 KHz sampling

rate resulted in frame time of 16 ms. The digitized speech data files were stored in a VAX 11/780 system where they were used by the processing algorithm. The DSC-200 was also used for playback of processed speech.

Software Development

All the software was developed on a VAX 11/780 running the VMS operating system. The program was developed in the FORTRAN-77 language. The program was developed in a modular fashion and each module was tested individually as it was developed. The program was executed on a microVAX III in order to reduce the run-time.

SPIRE (6:6-12)

SPIRE is a software package that allows the user to examine and process the speech and audio signals. It takes full advantage of the Symbolics 3600 LISP machine's built in graphical capabilities. It provides bit-mapped display which is either 1280 pixels wide by 760 pixels high or 1216 pixels wide by 773 pixels high. There are many types of displays available from SPIRE. However, only the following displays were used for analysis of speech files.

(a) Original (time) waveform

(b) Wide Band Spectrogram

(c) Narrow Band Spectrogram

Figure 2.1 shows an example of the above mentioned displays of a sample utterance, "Air to Surface". Figure 2.1 (b) shows the Wide Band Spectrogram of the utterance. SPIRE calculates this spectrogram with a bandwidth of 300 Hz. Figure 2.1 (c) is the Narrow Band Spectrogram, computed with a bandwidth of 78 Hz. In all three displays, there is a "marker" located at 0.5547 seconds of the original waveform. When the marker position is changed in any one display then it automatically changes its place in the other two displays.

ILS (8:1)

ILS is a software program that offers comprehensive software solution for signal processing. This software package works in conjunction with graphical terminals such as a VAX station. A wide range of signal processing functions are available in ILS. Signals may be displayed as numeric values or as waveforms. ILS was used only for obtaining displays at different points in the program. These displays were used for visual analysis of speech files before making rules for the speech processing program.

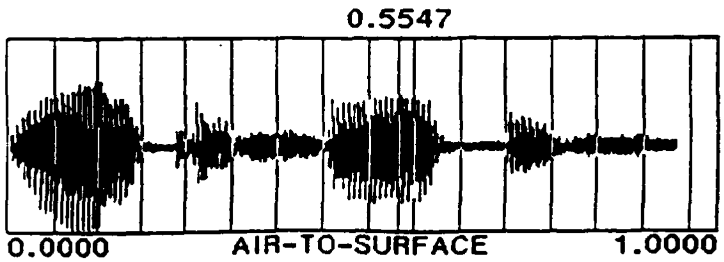


Fig. 2.1 (a) Time Waveform

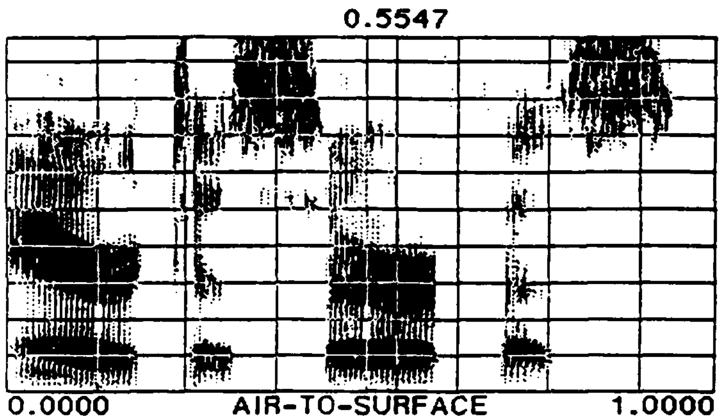


Fig. 2.1 (b) Wide Band Spectrogram

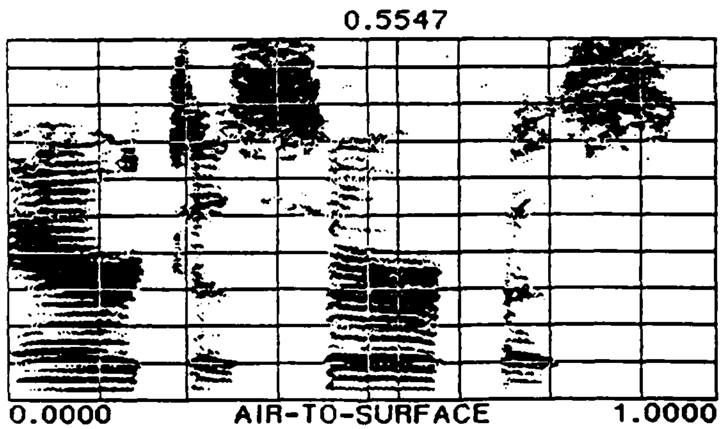


Fig. 2.1 (c) Narrow Band Spectrogram

III. Speech Processing System

Introduction

The purpose of this chapter is to describe the system design. This chapter will provide details about the major processing functions and how they are used. Figure 3.1 diagrams the modules of the processing system. Each module is described below.

Hamming Window

A Hamming window (equation 3.1) reduces the high

$$W(n) = 0.54 - 0.46 \operatorname{Cos}\left(\frac{2 \pi n}{256} \right) \quad \ldots\ldots\ldots \quad (3.1)$$

frequency ringing effects that would be caused by sampling speech with a rectangular window. As the sampled data has 256 points per frame, so the Hamming window also has 256 points in each frame. Thus each Hamming window covers 16 ms. If the window and hence the DFT are applied to nonoverlapping frames of speech data, a significant part of the data is ignored due to small values of window near the boundaries. Short-duration tone like signals near the boundaries can be missed (7:56). To avoid this loss of data, the window is usually applied to the overlapped frames. An overlapping scheme of 2:1 is used, as shown in

```
+------------------+   N frames    +----------------+
| Digitized Speech +-------------->| Hamming Window |
+------------------+               +-------+--------+
                                           |
                                           |  2N-1
                                           |  frames
                                           v
+--------------------+          +---------+--------+
| Spectrum Smoothing |<--------+   512-Point DFT    |
|                    |   Amp.  | (amplitude & phase)|
+---------+----------+          +---------+--------+
          |                               |
          v                               |
+---------+----------+                    |
| Amplification of   |                    |  Phase
| High Frequencies   |                    |
+---------+----------+                    |
          |                               |
          v                               v
+---------+-------+                  +----+-------+
| Harmonics/Peaks +----------------->|   Speech   |
|   Selection     |                  | Synthesis  |
+-----------------+                  +----+-------+
                                          |
                                          |
                                          v
+--------------+             +-----------+---------+
|  Amplitude   |<-----------+ Averaging the Number |
| Normalization|  N frames  |  of output frames    |
+-------+------+             +---------------------+
        |
        |
+-------+--------------------+
| Processed Digitized Speech |
+----------------------------+
```

Fig. 3.1 The Speech Processing System

figure 3.2. Thus each 16 ms frame begins 8 ms after the start of previous frame.

```
|          N frames :  256 points/frame              |
|____________________________________________________|
0 ms                                                 n ms

|  frame # 1  |
|_____________|
0 ms          16 ms

       |  frame # 2  |
       |_____________|
       8 ms          24 ms

                 --------------------------

                                |frame # 2N-1 |
                                |_____________|
                              (n-16) ms      n ms
```

Fig. 3.2   Illustration of Application of
           Hamming Window with 50% Overlap

If the original speech has N 256-point frames then this 50% overlap will produce (2N-1) 256-point frames.

Discrete Fourier Transform

A typical 512-point DFT routine (9:457) processes each 256 samples of the frame. In order to get 512-point DFT of 256 sample points of a frame, 256 zeros were added at the end of these sample points. The complex values were set to zero. The 512-point DFT produces 256 points of real and imaginary values of frequency spectrum. This gives a

resolution of 31.25 Hz in the spectrum. The magnitude and phase of the spectrum are calculated from the real and imaginary parts of the DFT. The phase values are stored to be used later for the synthesis of speech.

Smoothing of Spectrum

The frequency spectrum of voiced speech, which is not corrupted by noise, changes smoothly from frame to frame. However, if the speech is mutilated by noise then the transition from frame to frame may not be smooth and the change can be very erratic because of addition of noise. In order to reduce this erratic change the smoothing of spectrum was carried out. For smoothing of n^{th} frame, frames (n-1) and (n+1) were added to frame n, point by point, and the resultant values were divided by 3 to get new values for n^{th} frame. All the frames in the speech file were smoothed using this method.

Amplification of High Frequencies

All the SEU-processed files had almost no energy content above 2.5 KHz in the frequency spectrum. It appeared that SEU passes all the speech files through a low pass filter with a cut-off frequency at about 2 KHz. This SEU process had eliminated all the fricatives and high-order frequency terms from the speech. Figure 3.3 shows a narrow band spectrogram of a SEU-processed speech.

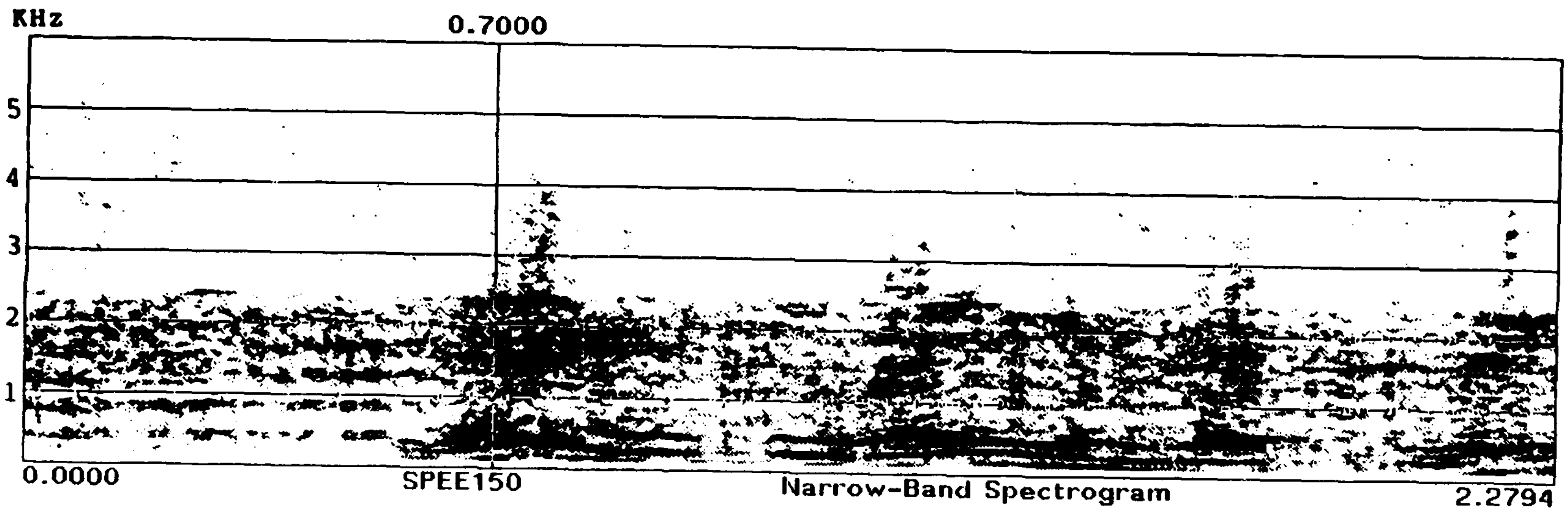


Fig. 3.3 Narrow Band Spectrogram

In order to increase the energy of the fricatives and high order harmonic frequency component of the glottal pulse, all the DFT points above 2.5 KHz were multiplied by a factor of ten. This multiplication will increase the energy in any noise components also, but that was mitigated to some extent when the harmonics/peaks were selected as described below.

Harmonics/Peaks Selection

The waveforms of voiced sounds are approximately periodic. The periodicity of a time waveform translates itself in the frequency domain as harmonics of fundamental frequency corresponding to the period of time waveform (10:4). Since the energy of the periodic signal is concentrated in bands of frequencies and the interfering signals, in general, have energy over the entire frequency band, the selection of harmonics of the fundamental frequency can eliminate the noise from the speech. The synthesis of voiced speech using the exact harmonics of the glottal frequency generated a " musical noise " in the synthesized speech. In order to eliminate this effect, values selected were not necessarily the exact harmonics of the glottal frequency. The two neighboring frequency points were checked and if any of those points had higher amplitude than the exact glottal harmonic then that frequency was selected instead of the harmonic.

The glottal frequency in human voice ranges from 100 Hz to 150 Hz for males and from 150 Hz to 250 Hz for females (13). All the SEU processed files were of male speakers. For this reason 125 Hz was selected as basic glottal frequency. This covered a frequency range from 93.75 Hz to 156.25 Hz, as the two adjacent frequency points were also checked for maximum value. Figure 3.4 depicts this process of harmonic selection in voiced regions of speech.

All the harmonics below a certain threshold of amplitude (a) were eliminated. This threshold varied for each speech file processed. Appendix A gives the value of this threshold for each file processed. This threshold was varied within the frame also because the frequency spectrum of voiced speech falls off at a rate of 6 dB per Octave after about 600 Hz. Equation 3.2 shows the threshold value as varied within the frame.

$$\text{Threshold} = \begin{cases} a & 0 < f < 600\ \text{Hz} \\ a\left(\dfrac{8150 - f}{240}\right) & f \geqslant 600\ \text{Hz} \end{cases} \quad \ldots\ (3.2)$$

Harmonic selection does not apply to the unvoiced speech because frequency components of unvoiced speech are not harmonically related to each other. For this reason a rule based on the energy of a frame was established to decide if a given frame contained voiced speech or not.

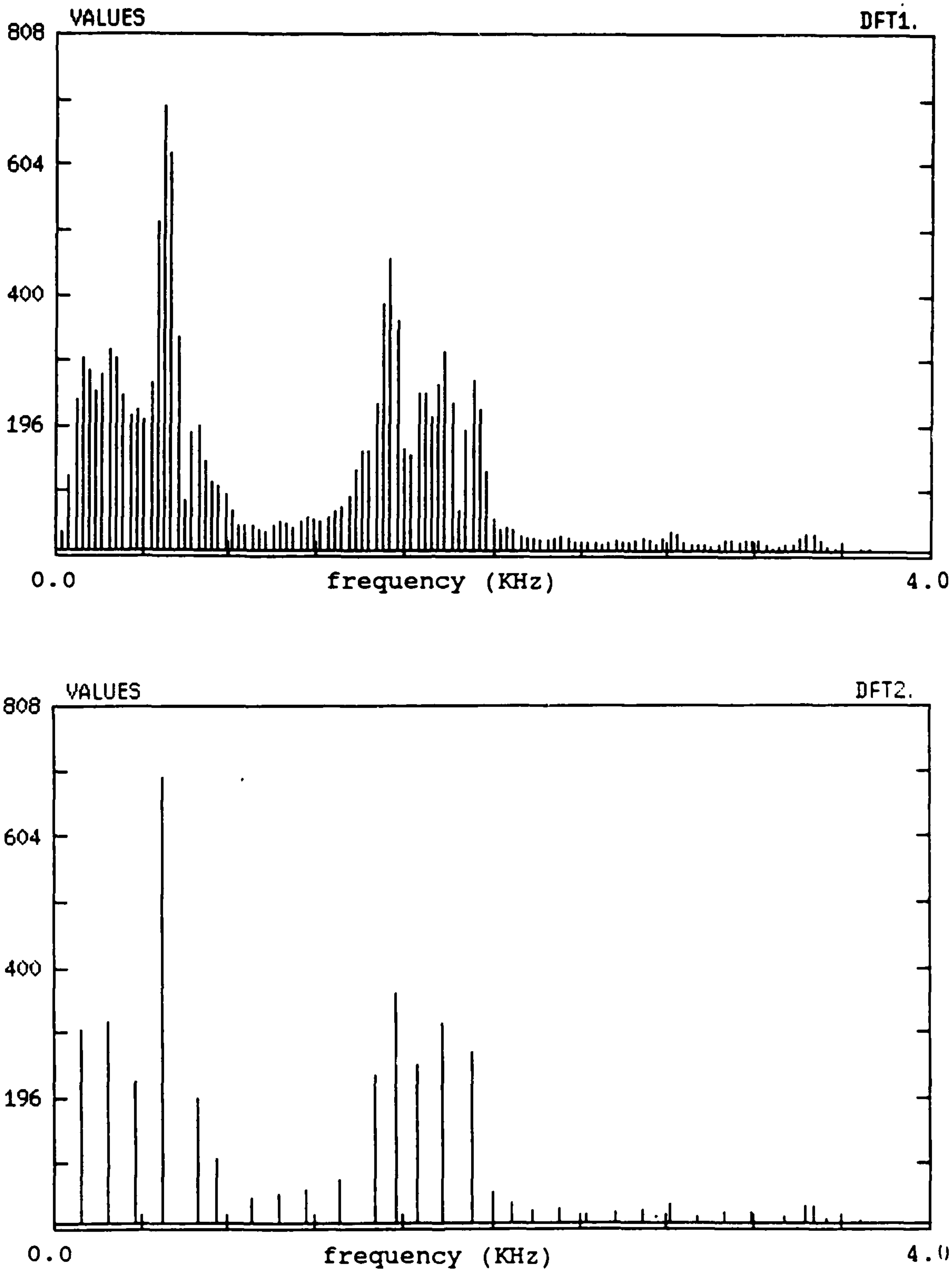


Fig. 3.4 Harmonics Selection

The energy of each frame was computed and checked against threshold. If the energy was below the threshold then the frame was considered to have unvoiced speech data. This energy threshold was empirically determined and fixed at 5.0×10^5 for all speech files. Once it was determined that the frame contained unvoiced speech data then only the peaks of spectrum in that frame were selected in order to minimize the noise. All the peaks below a certain threshold were eliminated. This threshold also varied for each speech file processed Appendix A gives the value of this threshold for each speech file. Figure 3.5 shows this process of selection of peaks in an unvoiced region of speech.

Speech Synthesis

Speech was reconstructed after processing the amplitudes of frequency spectrum of speech. This reconstruction was based on the fact that speech can be represented as a sinusoidal model (12:489). The modified amplitude, frequency, and the original phase were used to reconstruct the speech. The formula for reconstruction of speech is given in equation 3.3 where S(n) is the reconstructed speech (in discrete time). The estimates of

$$S(n) = \sum_{n=1}^{256} \sum_{i=1}^{256} amp_i \cos (2 \quad f_i \, t_n + phs_i) \quad ..(3.3)$$

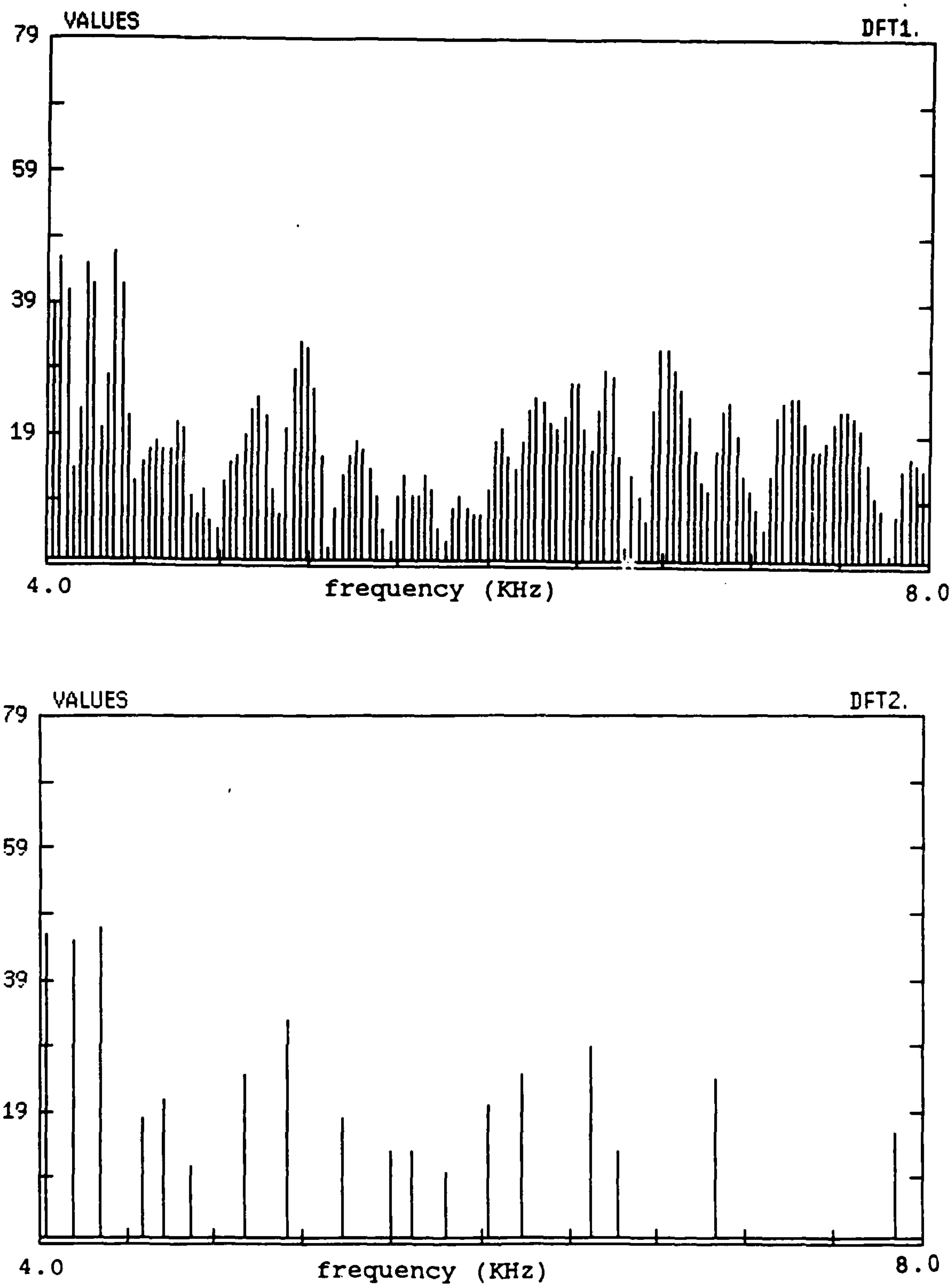
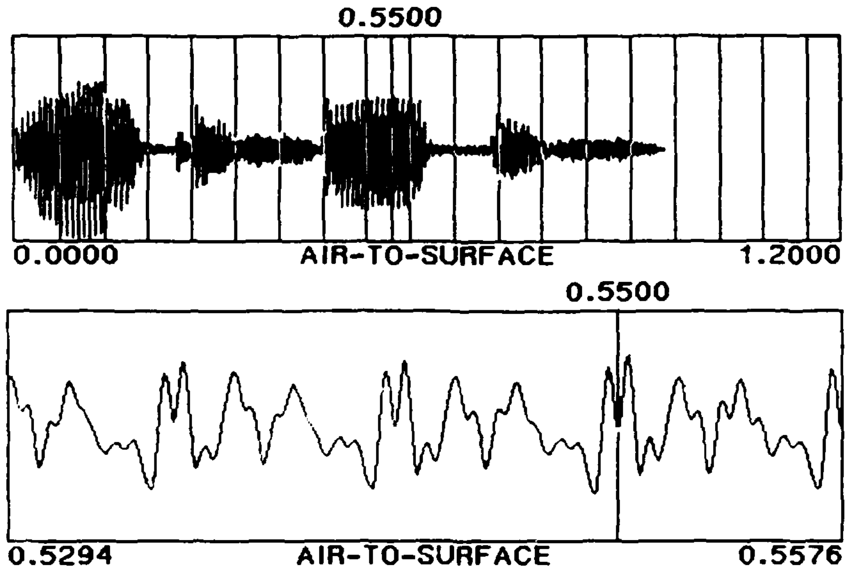


Fig. 3.5 Selection of Peaks

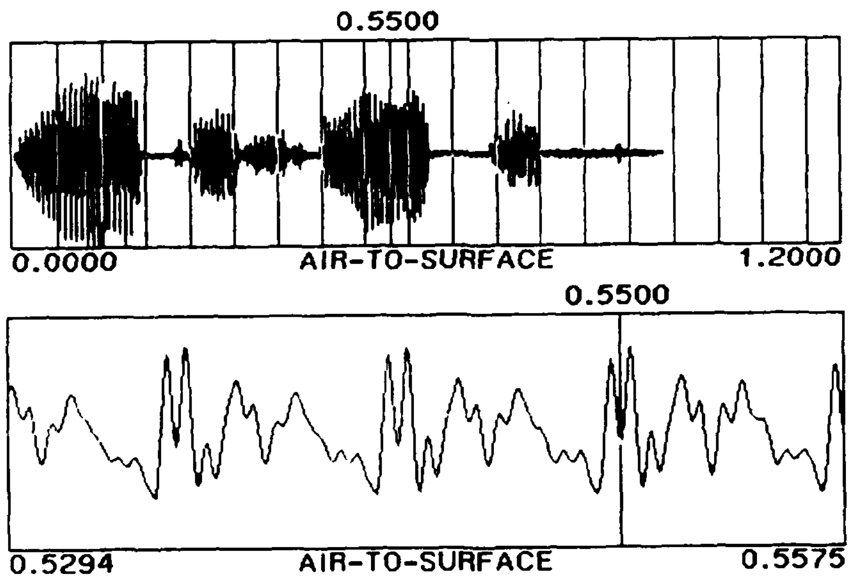
parameters which are used in generating s(n) correspond to the modified amplitude (amp), frequency (f), and phase (phs) that were measured from the DFT at the location of harmonics/peaks selected. If the original speech is free of noise then the resulting synthesized waveform preserves the original waveform shape and is essentially perceptually indistinguishable from the original speech. Figure 3.6 shows the time waveform of a lab-recorded utterance and the synthesized utterance. The narrow band spectrogram of these utterances is shown in figure 3.7.

Averaging the Number of Frames

Once the Hamming window with 50% overlap was applied to the original speech, the number of speech frames had increased from N to 2N-1. In order to reduce the number of frames to original number N, the averaging of the frames was carried out. The odd numbered frames correspond to the original frames whereas the even numbered frames were the result of 50% overlap. To reduce the number of frames and to eliminate any discontinuity in splicing the frames together, Hamming window was applied to all 2N-1 frames. After application of the Hamming window, the first half of an even numbered frame was added to the second half of previous frame, point by point, and second half of that even numbered frame was added to the

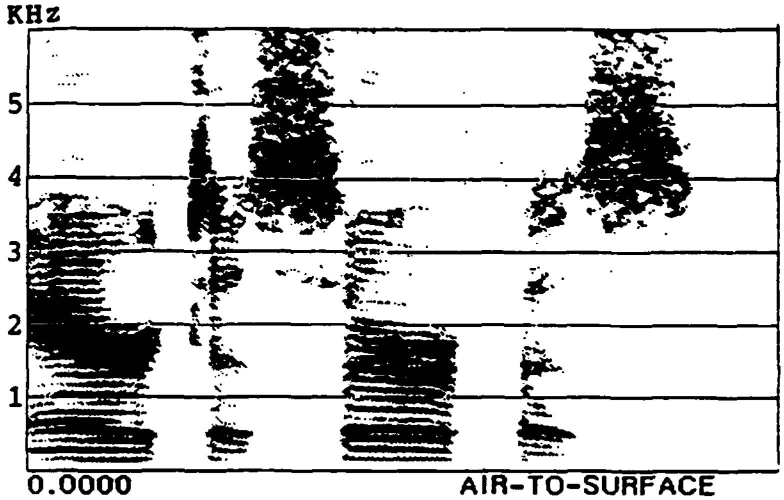


Original

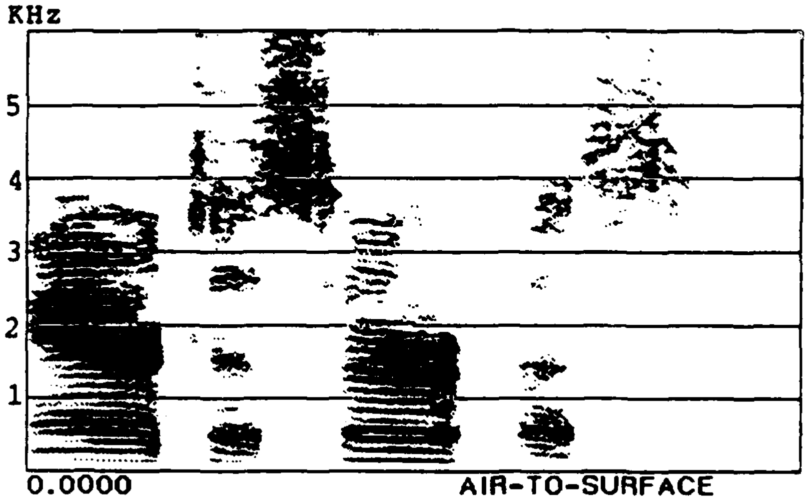


Synthesized

Fig. 3.6 Time Waveforms



Original



Synthesized

Fig. 3.7 Narrow Band Spectrogram

first half of the next frame. The even numbered frames were then removed to reduce the number of frames to N.

Amplitude Normalization

The DSC-200 accepts only integer*2 data type for conversion to analog speech waveform. The synthesized speech was in real format. In order to change it to integer*2 format, the absolute maximum amplitude in the synthesized speech was estimated and all the sample points in the speech were divided by this maximum value and then multiplied by 32767. Multiplication by 32767 was carried out because integer*2 data type can have a maximum value of 32767. By this amplitude normalization all the processed speech files had same level of volume irrespective of their input levels.

IV. Results and Discussion

Introduction

The purpose of this chapter is to examine the effect of different modules of the speech processing algorithm on SEU processed speech and to present the overall results of the speech processing algorithm on SEU processed speech.

Smoothing of DFT

The smoothing of frequency spectrum was employed in order to reduce the erratic changes of amplitude spectrum from frame to frame and to minimize the effect of noise on the voiced regions of speech. Two different orders of smoothing were tried. First, for smoothing of n^{th} frame, $1/4^{th}$ amplitude of frequency components of frame number (n-1) and (n+1) were added to $3/4^{th}$ amplitude of frequency components of frame number n, point by point, to get the new amplitude values for n^{th} frame. This method, however, did not produce acceptable results. Secondly a more rigid smoothing was tried. In this case, amplitudes of frequency components of all three frames were added together, point by point, and the resultant values were divided by 3 to get new amplitude values for n^{th} frame. This smoothing scheme provided better quality speech than the other. Figure 4.1 shows three consecutive frames of DFT without

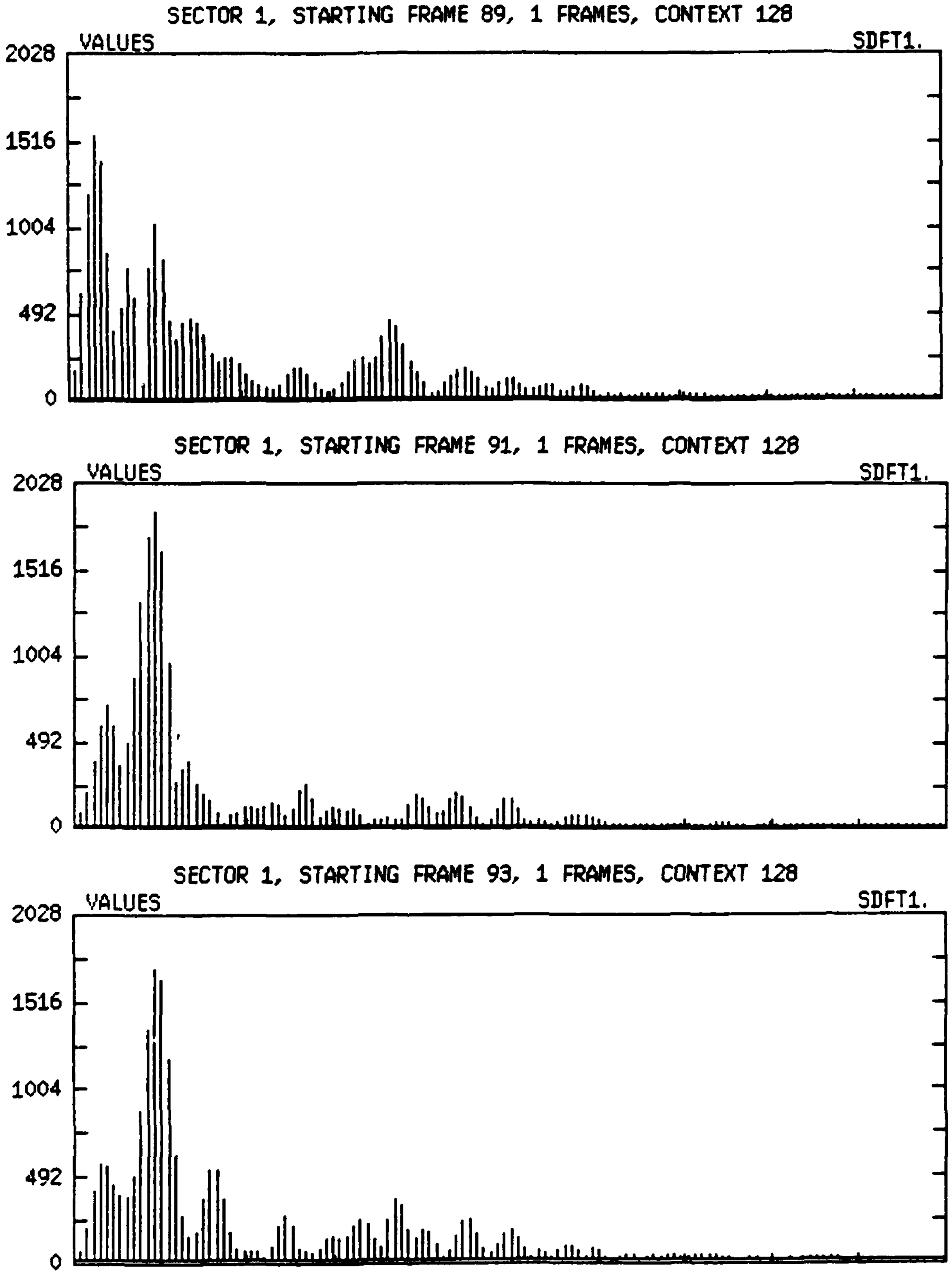


Fig. 4.1 Amplitude Spectrum without Smoothing

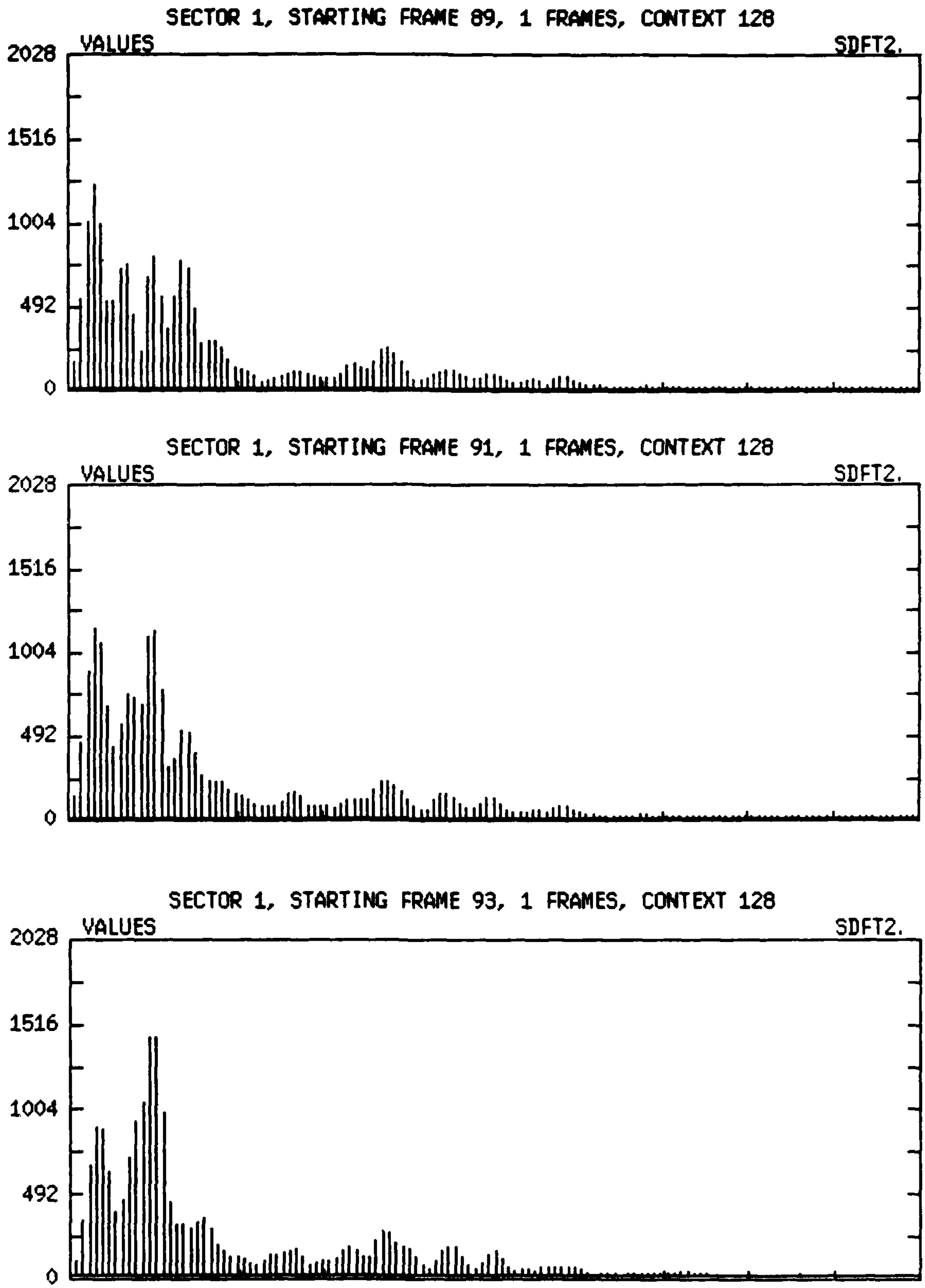
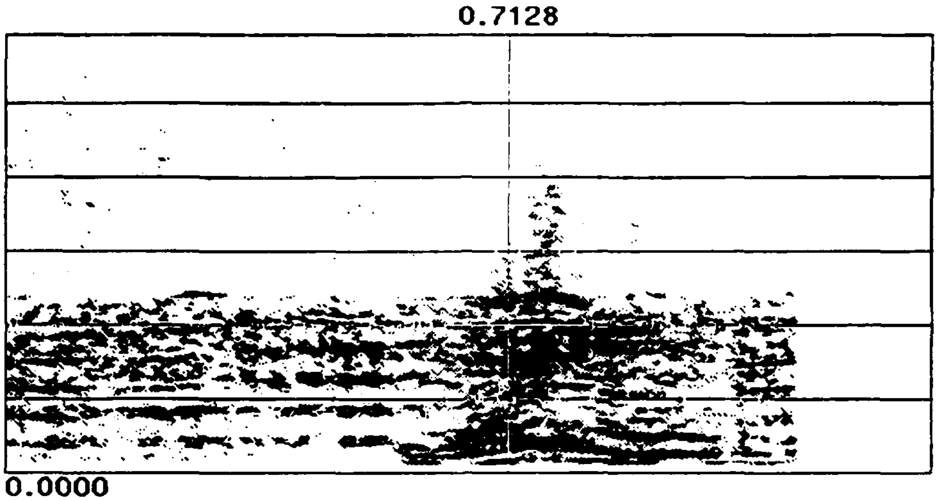


Fig. 4.2 Amplitude Spectrum with Smoothing

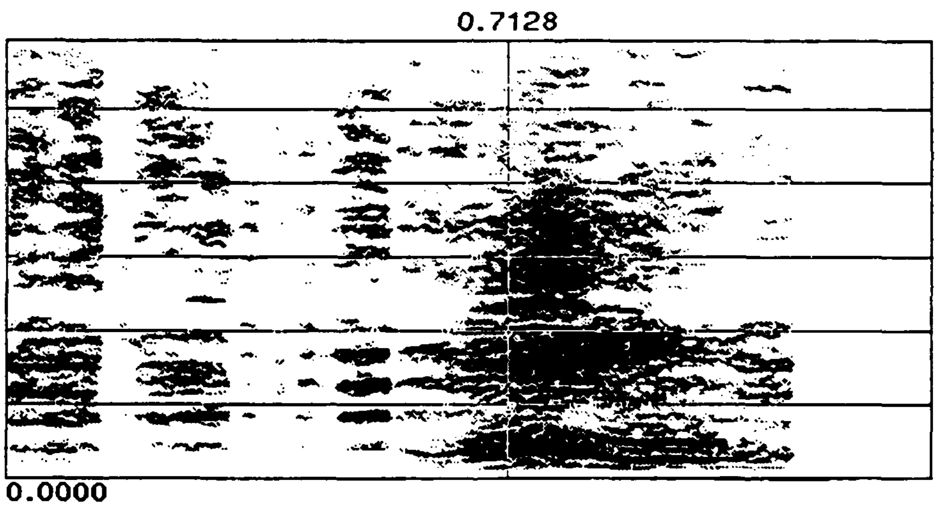
and figure 4.2 shows the effect of smoothing on the same three frames. Frame to frame variation of amplitudes of frequency components in figure 4.2 is much more smooth than that in figure 4.1.

High Frequency Enhancement

In speech the frequency content of fricatives have high energy above about 3 KHz and relatively very low energy below 3 KHz. If speech is passed through a low pass filter with a cut off at 2.5 KHz, the fricatives are attenuated drastically and the resulting mutilation of speech reduces the speech quality significantly. This was a prime reason for the reduced quality of SEU processed speech. In order to enhance the high frequency components, the amplitude spectrum above 2.5 KHz was amplified. Different amplification factors were tried. The best results were achieved once the frequency components above 2.5 KHz in voiced regions (high energy frames) were amplified by a factor of 10 and those in unvoiced region (low energy frames) were amplified by a factor of 5. The quality of speech improved after amplification of high frequency amplitudes. Figure 4.3 shows the narrow band spectrogram of a portion of SEU processed speech and the spectrogram of the same portion after high frequency enhancement.



(a) Original (SEU processed)



(b) High Frequencies Enhanced

Fig. 4.3 Narrow Band Spectrogram

Noise Cancellation

To improve the signal-to-noise ratio (SNR) of speech corrupted by broad band noise, the spectral noise subtraction method is often used (4;5). In this method the noise spectrum is estimated from speech, based on the low energy frames, and this noise estimate is subtracted from the speech spectrum, setting negative values to either zero or to a preset minimum level. This method improves the SNR considerably. This noise cancellation procedure was used for SEU processed speech also. The amount of noise in the speech reduced considerably but so did the intelligibility of the speech. The amount of annoying " musical " noise introduced by this process was not considered worth the improvement in the SNR. This noise cancellation process was eliminated as the main objective of this work was to increase the quality of the speech.

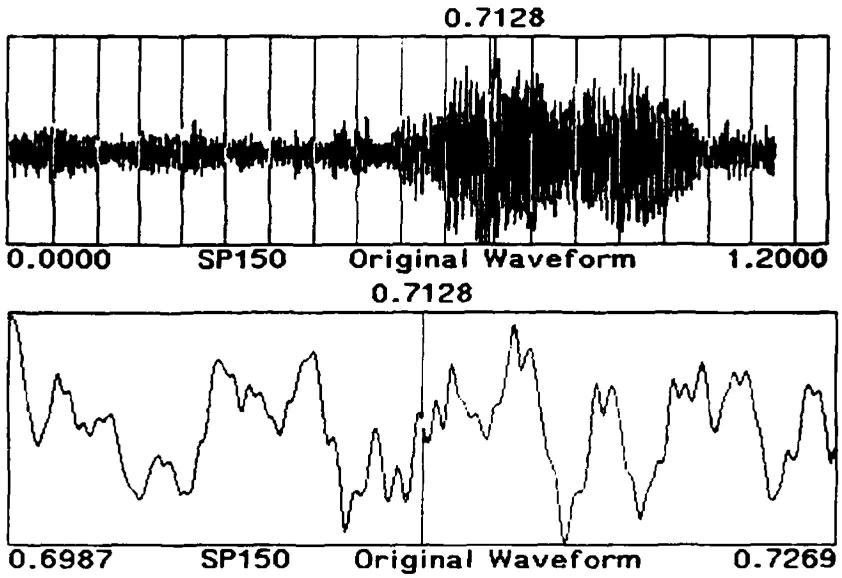
Harmonics/Peaks Selection

As the energy of the voiced speech is concentrated in bands of frequencies, selection of these bands helped eliminate the unnecessary noisy components in the spectrum. However, selection of exact harmonics was avoided as it introduced the " musical " noise in speech. Selection of frequency components by monitoring the two neighboring frequency amplitudes of the exact harmonic for maximum amplitude did not introduce the musical noise.

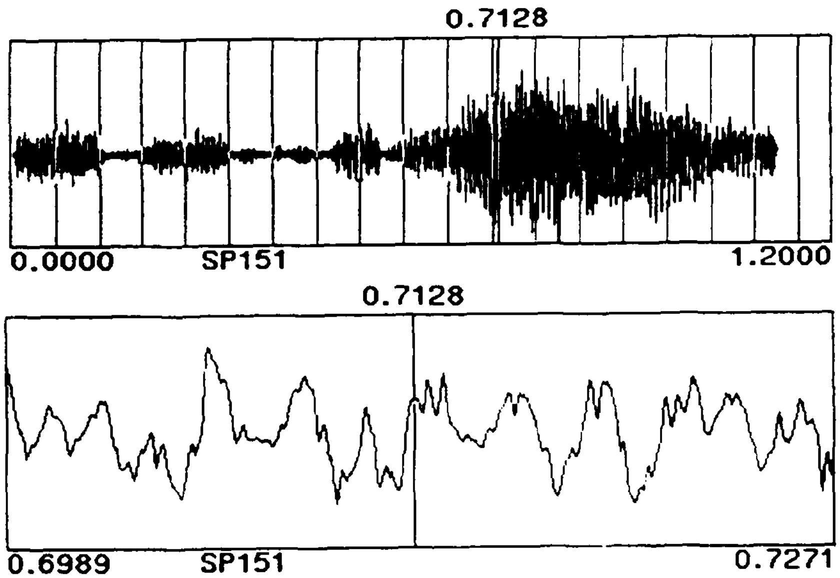
After selection of maximum value from the exact glottal frequency and its two neighboring components, the values selected were compared against an amplitude threshold. If the selected components were below that threshold then they were set equal to zero. This helped in reduction of noise without compromising the quality of the speech. This also reduced the computation time for resynthesis of speech. The threshold varied for different speech files and the values are given in Appendix A for all the speech files processed.

Speech Synthesis

The results of reconstruction of speech based on the sinusoidal model were very encouraging. The result of reconstruction of a sample utterance, free of noise, was perceptually indistinguishable from the original speech. The results of this processing algorithm for SEU processed speech were also encouraging. The overall quality of the speech improved considerably. Figure 4.4 compares the time waveform of a SEU processed speech and the reconstructed speech after this process. The narrow band spectrogram is compared in figure 4.5.

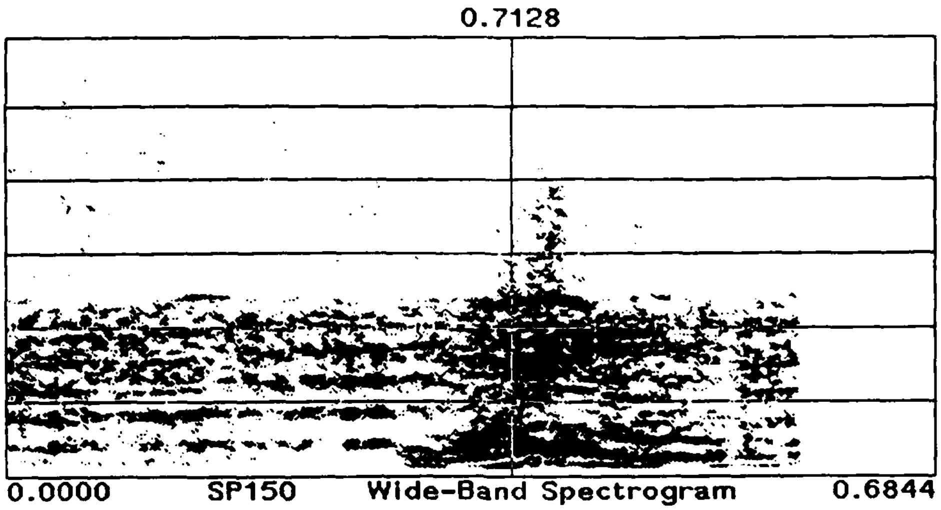


(a) SEU Processed Speech

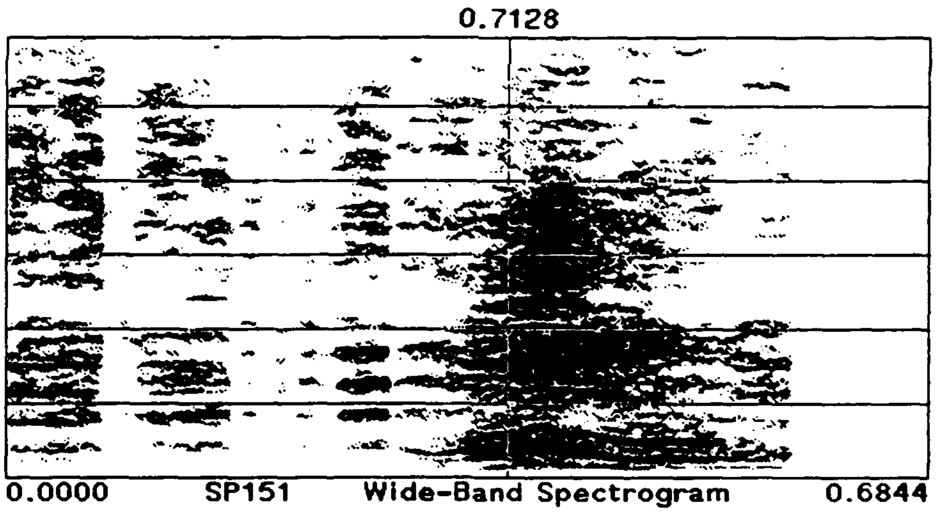


(b) Reconstructed Speech

Fig. 4.4 Time Waveform



(a) SEU Processed Speech



(b) Reconstructed Speech

Fig. 4.5 Narrow Band Spectrogram

V. Conclusions and Recommendations

Introduction

The purpose of this chapter is to discuss conclusions that may be drawn based on the performance of this system as well as to give recommendations for further research in this area of post-processing of speech.

Conclusions

This thesis is successful in producing a system that increases the quality of the SEU processed speech appreciably. The smoothing of amplitude spectrum of voiced regions of speech reduces the effects of additive noise on the speech spectrum. Results show that enhancement of high frequency components of amplitude spectrum of a filtered (low pass) speech can improve the quality of speech. The idea of harmonic selection by monitoring the two neighboring frequency components for maximum amplitude was also shown to be advantageous. The reconstruction of speech using a sinusoidal model works well.

Recommendations

Further investigation in the area of noise cancellation can further improve the results. Assuming the noise is stationary in a speech file, the estimate of average noise was quite accurate. However the subtraction

of this noise did not provide acceptable results as far as quality of the speech is concerned. The noise cancellation process can be investigated further to improve the results.

The use of harmonic selection and reconstruction of speech using a sinusoidal model may help in speaker independent speech recognition system. The glottal pitch frequency and the selected harmonics from a speech of a speaker can be translated in frequency to coincide with the glottal pitch frequency of the speech used for the template. This may produce better recognition results.

Summary

In summary, this thesis shows that using a sinusoidal model for speech and selection of harmonics can be successfully applied to the problem of SEU processed speech. Presumably, further improvements could be made by careful cancellation of noise spectrum from the speech spectrum. Consequently, further research in this area could help to ultimately solve the problem of correction of mutilated speech.

Appendix A: Sample Results

The results of this processing on different SEU Processed speech files are given in this appendix. The results are compared in the form of Time Waveform and Narrow Band Spectrogram. As mentioned earlier, threshold values varied from speech file to speech file. These values are also given in this appendix.

Figure A.1-1

SPEE100 SEU Processed Speech

Figure A.1-2

SPEE101 Reconstructed Speech

Amplitude threshold for harmonics : 30000

Amplitude threshold for peaks : 2000

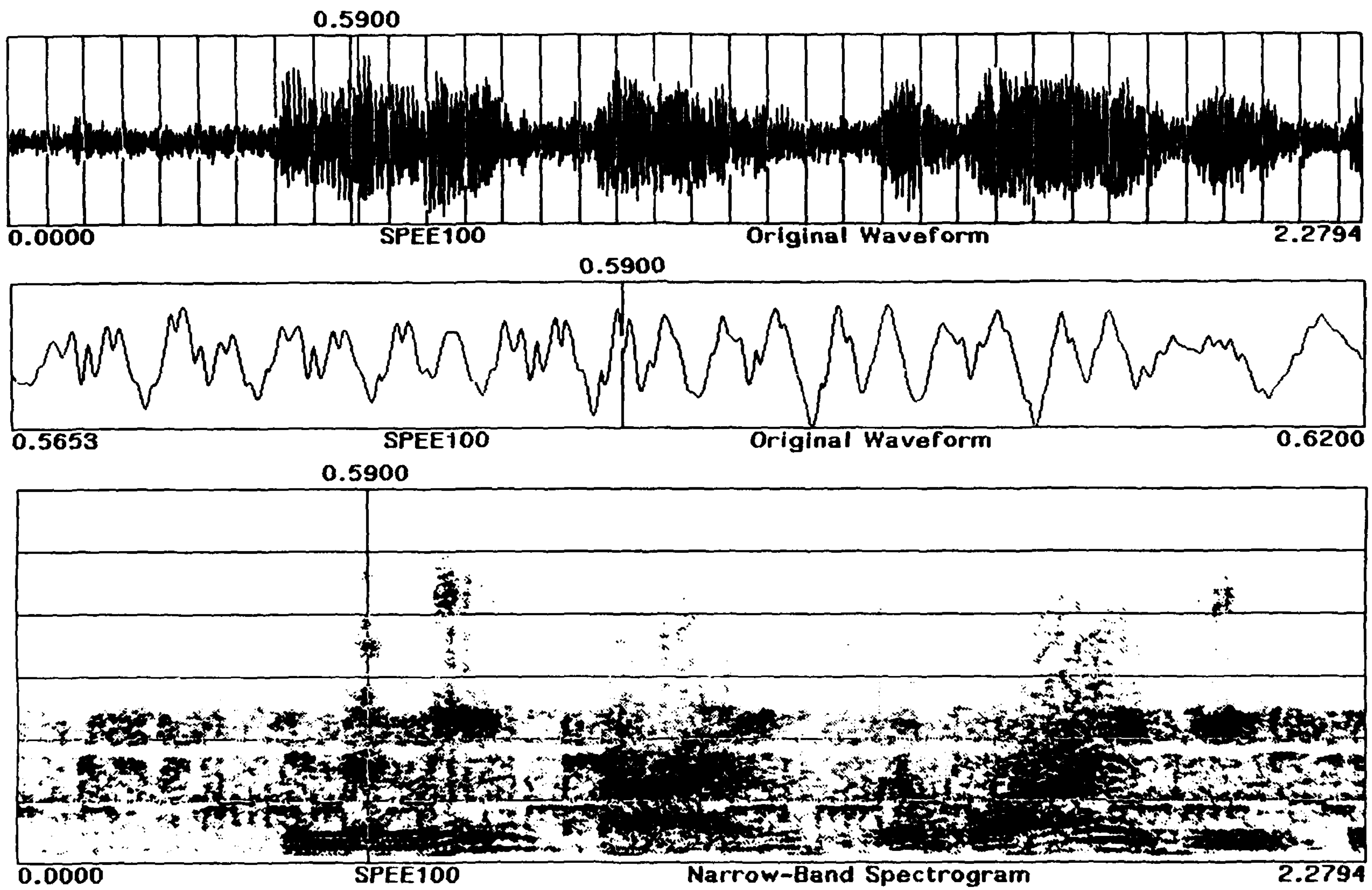


Fig. A.1-1 SEU Processed Speech

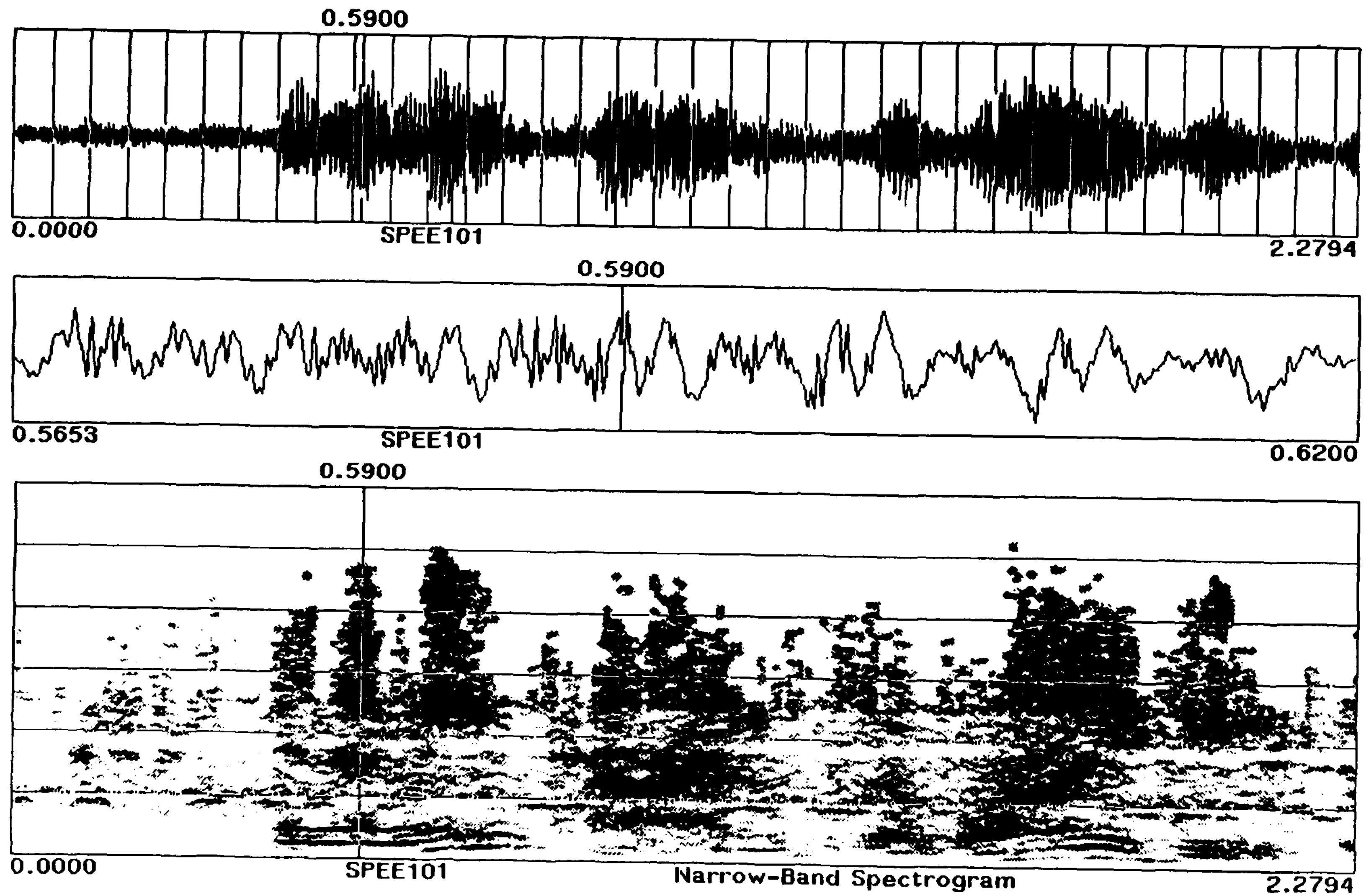


Fig. A.1-2 Reconstructed Speech

Figure A.2-1

SPEE150	SEU Processed Speech

Figure A.2-2

SPEE151	Reconstructed Speech

Amplitude threshold for harmonics	:	60000
Amplitude threshold for peaks	:	3000

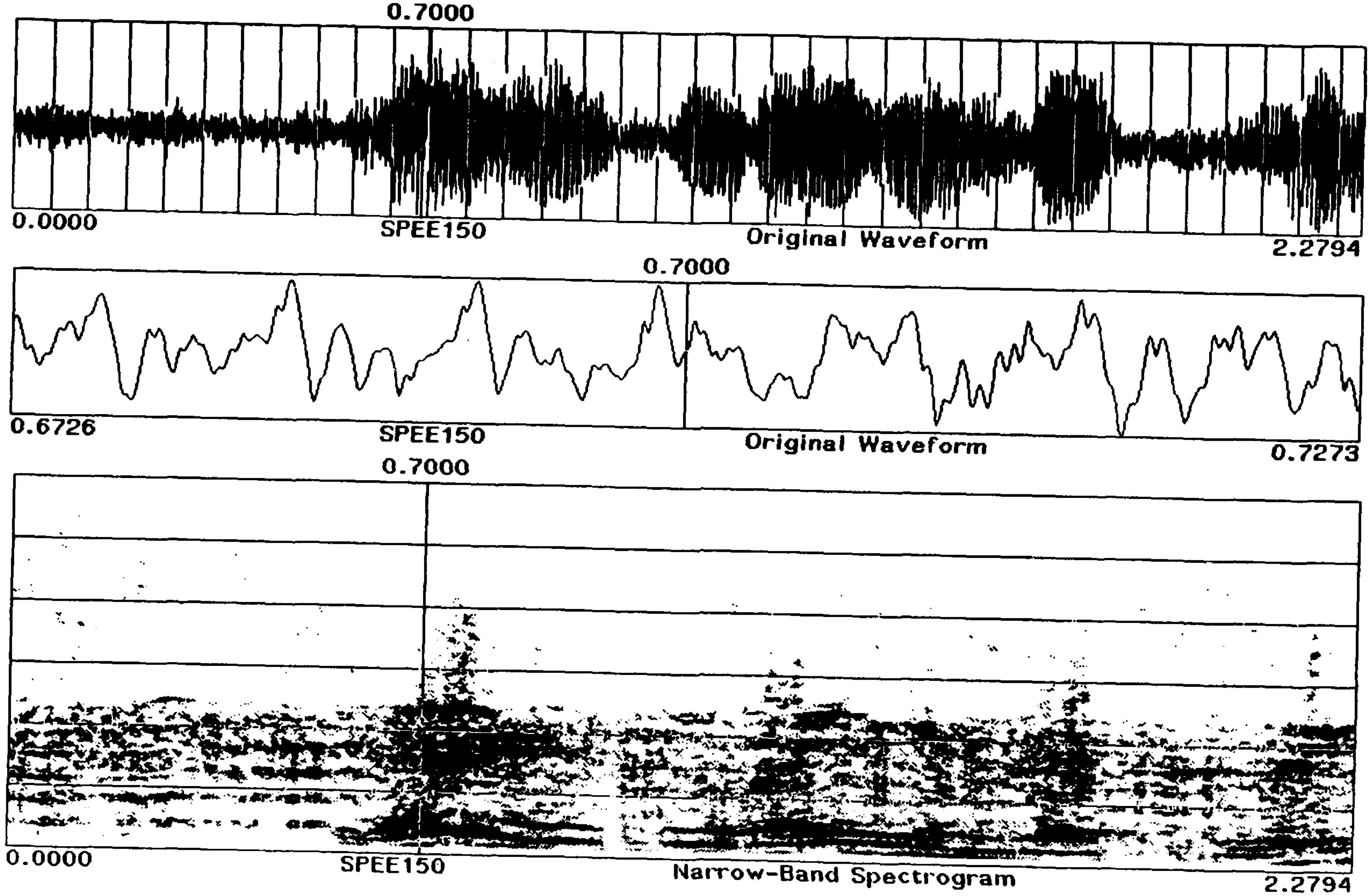


Fig. A.2-1 SEU Processed Speech

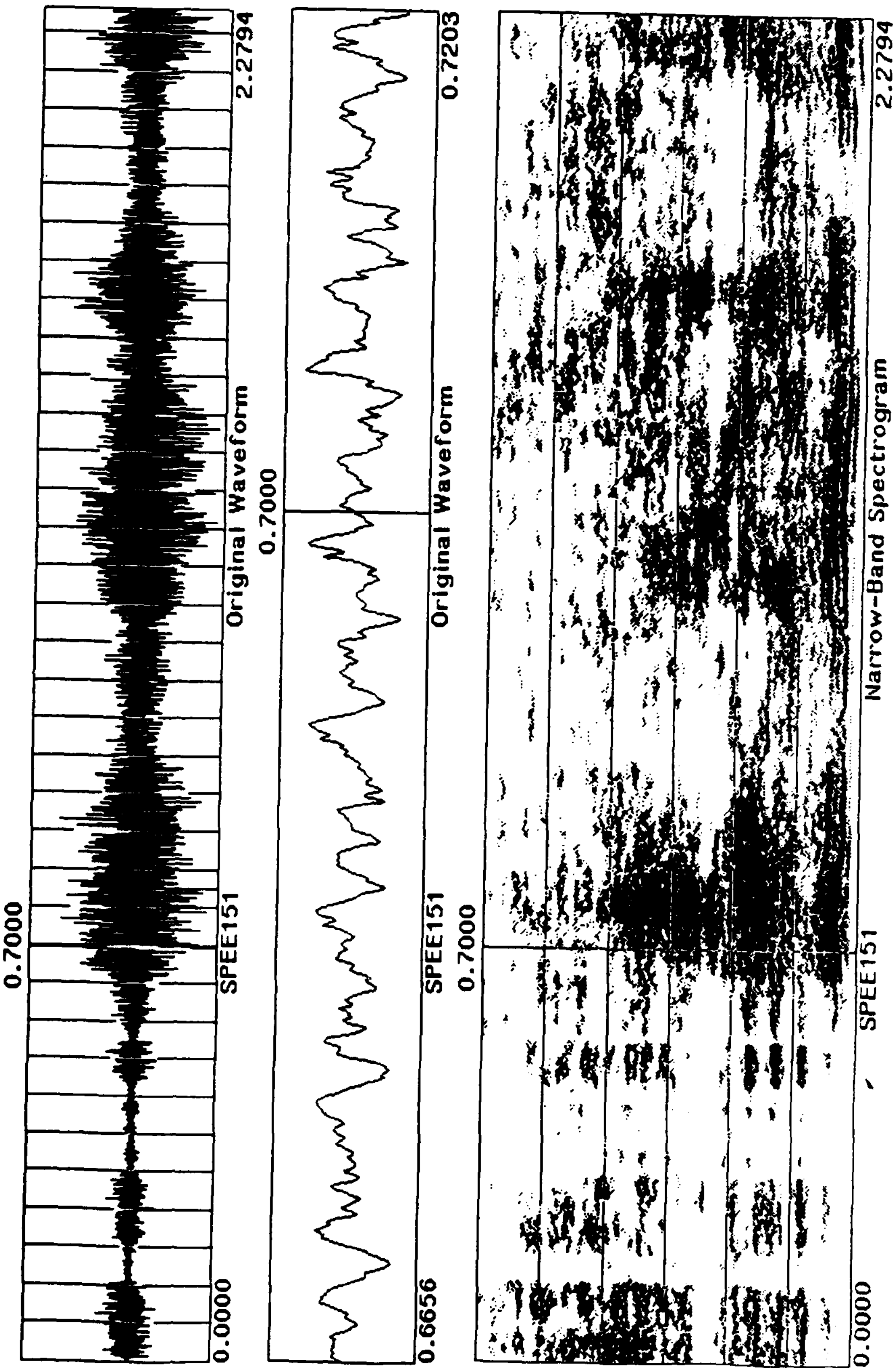


Fig. A.2-2 Reconstructed Speech

Figure A.3-1

SPEE200	SEU Processed Speech

Figure A.3-2

SPEE201	Reconstructed Speech

Amplitude threshold for harmonics : 10000

Amplitude threshold for peaks : 1000

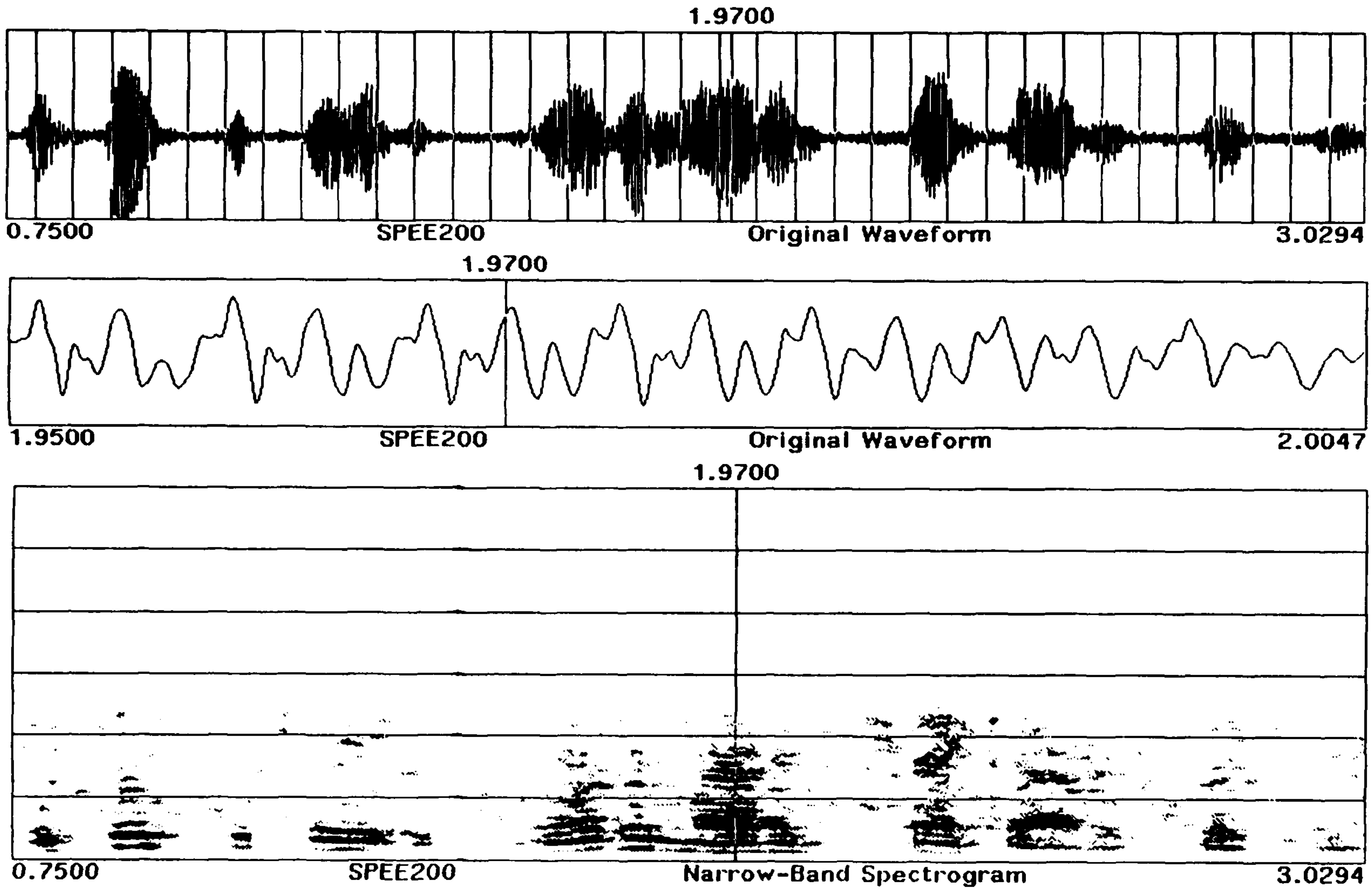


Fig. A.3-1 SEU Processed Speech

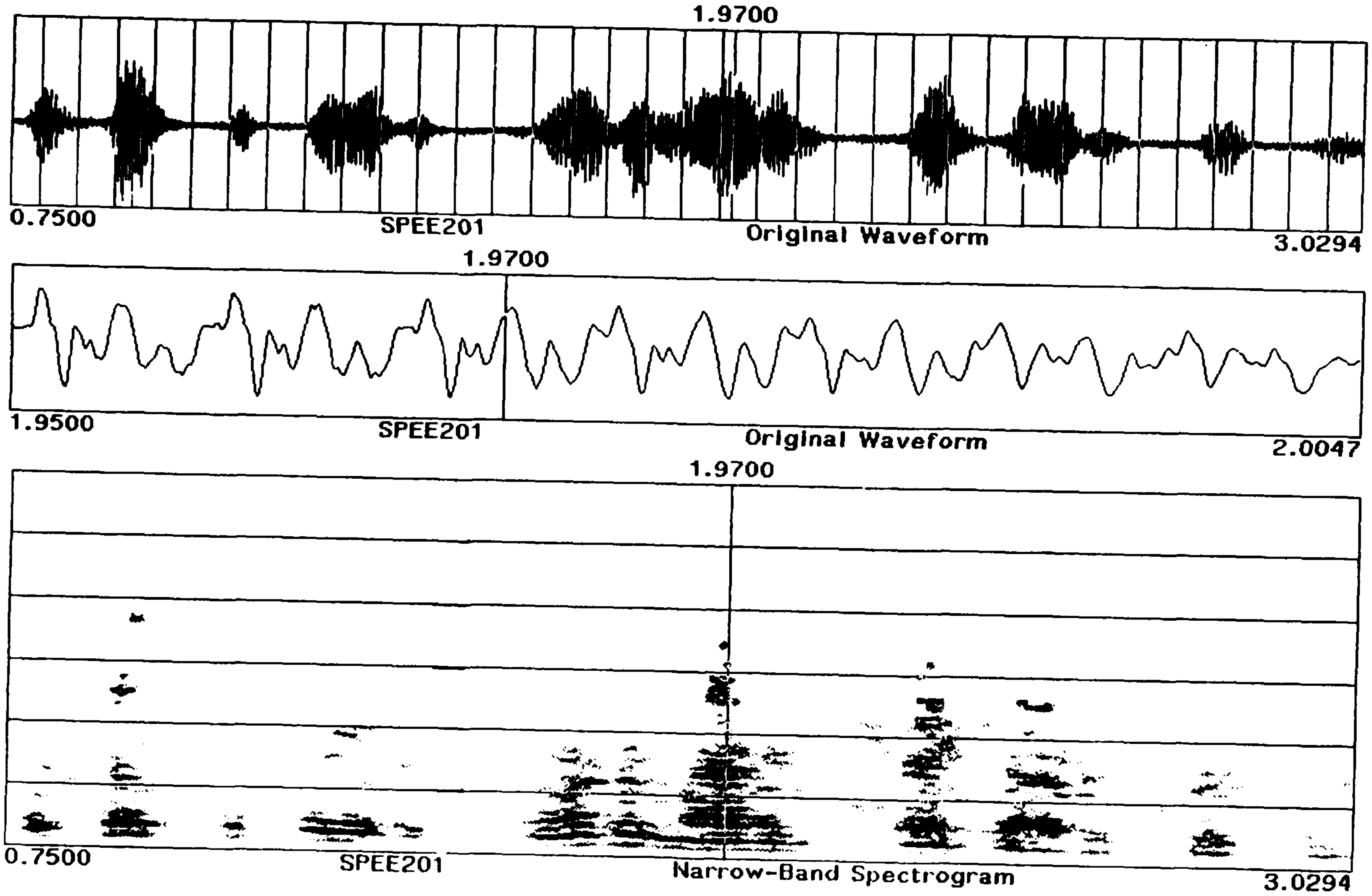


Fig. A.3-2 Reconstructed Speech

Figure A.4-1

SPEE250	SEU Processed Speech

Figure A.4-2

SPEE251	Reconstructed Speech

Amplitude threshold for harmonics : 35000

Amplitude threshold for peaks : 2000

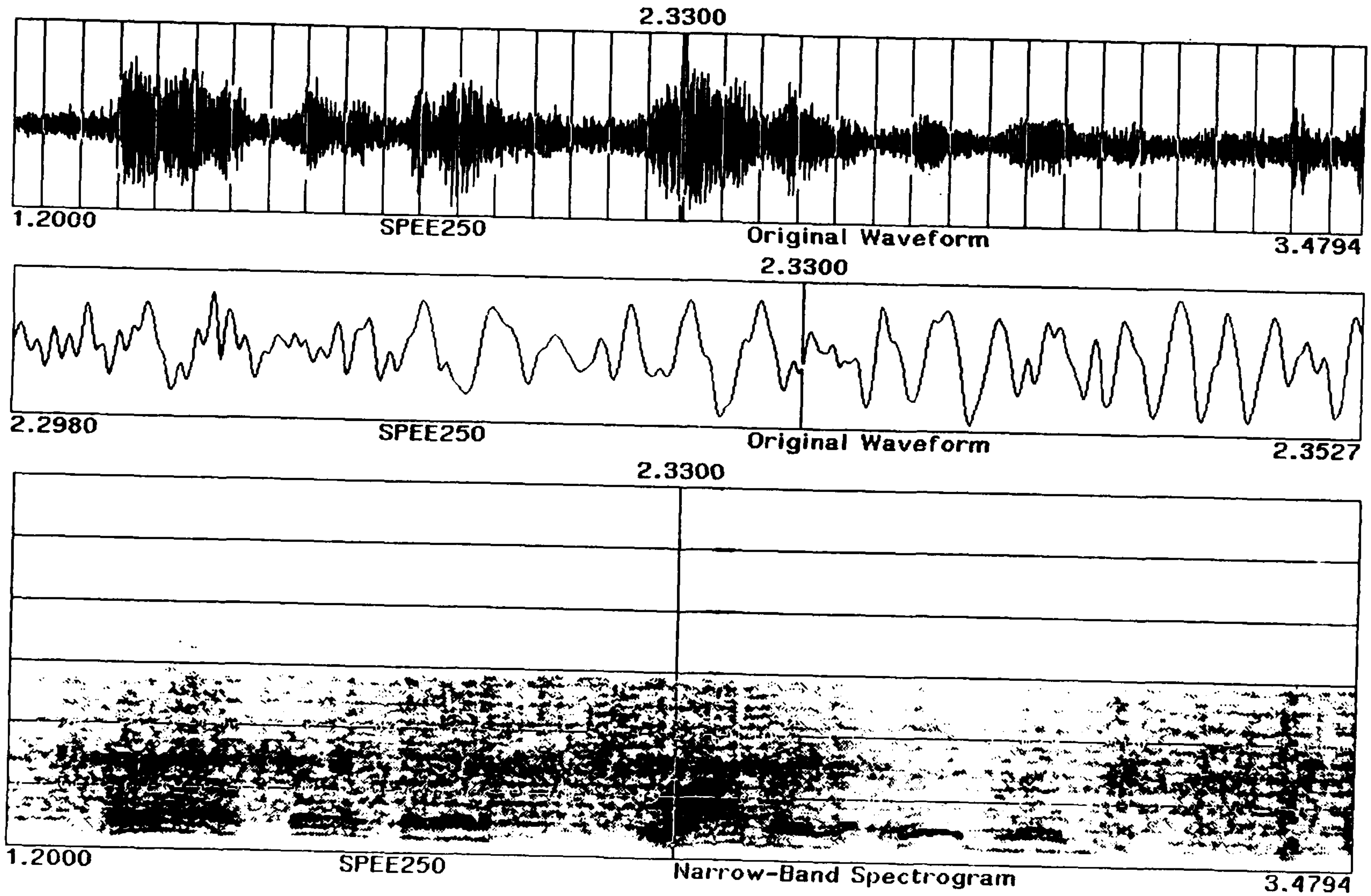


Fig. A.4-1 SEU Processed Speech

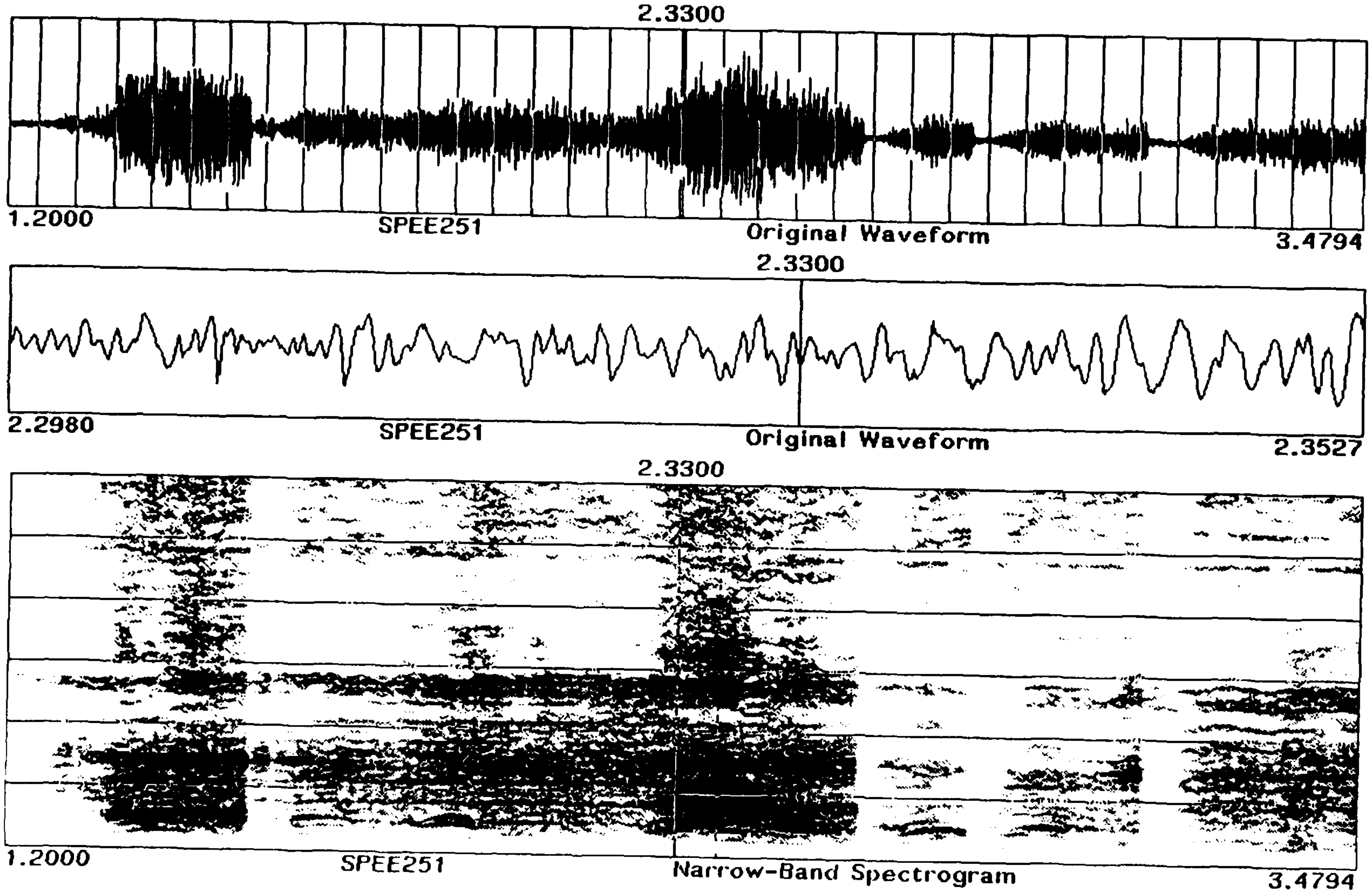


Fig. A.4-2 Reconstructed Speech

Appendix B: Program Listing

```
c*************************************************************
c                                                            *
c   Title :  Speech Processing                              *
c   Author:  Flt.Lt. Naddem A. Bashir                        *
c   Date  :  February 1989                                   *
c                                                            *
c   Function:                                                *
c            This program processes a digitized speech file*
c            and improves its quality and intelligibility. *
c            It needs the name of input speech file and the*
c            name to output the processed speech.            *
c                                                            *
c   Environment:                                             *
c            This is a Fortran 77 program that has been      *
c            designed to run on a VAX 11/780 machine.        *
c                                                            *
c*************************************************************

      dimension u(256),v(256),w(256),x(256),y(256),z(256)
      dimension pk(256),ph(256),rdata(256),xx(512),yy(512)
      real x,y,a,b,rdata,pi,z,u,pk,amp,v,w
      real amax,ph,phas,xx,yy,c,p,amax1,noise,thresh
      real en1,en2,alpha,beta
      integer i,j,k,n,t
      integer*4 hdata(64)
      integer*2 idata(256),s(256)
      character*32 in,out
      data pi/3.14159265358979324/
      write(*,4)'          Enter the Name of Input File:'
      read(*,6) in
      write(*,4)'          Enter the Name of output File:'
      read(*,6) out
4     format(a,$)
6     format(a)
9     open(10,file=in,access='sequential',status='old',
     +     recordtype='fixed',form='unformatted',recl=128)
      open(99,file=out,access='sequential',status='new',
     +     recordtype='fixed',form='unformatted',recl=128)
      read(10) hdata
      write(99) hdata
c*************************************************************
c                                                            *
c   Applying the Hamming window to the input data with 50% *
c   overlap. The number of frames is increased to (2N-1),  *
c   where N is the number of original frames.               *
c                                                            *
c*************************************************************
      print*,' APPLYING HAMMMING WINDOW WITH 50% OVERLAP '

      open(30,status='scratch',form='unformatted',
     +        recordtype='fixed',recl=256)
      j=1
100   read(10,end=199) idata
      k=mod(j,2)
```

```
      if(k.eq.0) goto 120
110   do i=1,256
         x(i)=idata(i)
         w(i)=x(i)*(0.54-0.46*cos((2*pi/255)*(i-1)))
      end do
      write(30) w
      j=j+1
      goto 100
120   do i=1,128
         z(i)=x(i+128)
         z(i+128)=idata(i)
      end do
      do i=1,256
         v(i)=z(i)*(0.54-0.46*cos((2*pi/255)*(i-1)))
      end do
      write(30) v
      j=j+1
      goto 110
199   rewind(30)
c*************************************************************
c                                                            *
c   512-Point Discrete Fourier Transform                     *
c                                                            *
c*************************************************************
      print*,' TAKING 512-POINT DFT ( PHASE AND AMPLITUDE )'
      open(40,status='scratch',form='unformatted',
     +        recordtype='fixed',recl=256)
      open(50,status='scratch',form='unformatted',
     +        recordtype='fixed',recl=256)
      do i=1,512
         yy(i)=0.
      end do
200   read(30,end=299)x
      do i=1,256
         xx(i)=x(i)
         xx(i+256)=0.
      end do
      call fft(9,xx,yy,1)
      do i=1,256
         a=xx(i)
         b=yy(i)
         x(i)=(sqrt(a**2+b**2)*1.7)
         if(a.eq.0) then
            y(i)=pi/2
            else
            y(i)=atan2(b,a)
         end if
      end do
      write(40)x
      write(50)y
      do i=1,256
         x(i)=0.
         y(i)=0.
```

```
      end do
      do i=1,512
         xx(i)=0.
         yy(i)=0.
      end do
      a=0.
      b=0.
      goto 200
299   rewind(40)
      rewind(50)
      close(30)
c***********************************************************
c                                                          *
c   Smoothing the DFT Amplitude                            *
c                                                          *
c***********************************************************
      print*,' SMOOTHING OF SPECTRUM '
      open(60,status='scratch',form='unformatted',
     +        recordtype='fixed',recl=256)
      j=1
300   read(40,end=399) rdata
      if(j.gt.1) goto 310
      do i=1,256
         x(i)=rdata(i)
         rdata(i)=0.
      end do
      write(60) x
      j=j+1
      goto 300
310   k=mod(j,3)
      if(k.eq.0) goto 320
      do i=1,256
         y(i)=rdata(i)
         rdata(i)=0.
      end do
      j=j+1
      goto 300
320   do i=1,256
         z(i)=(x(i)+y(i)+rdata(i))/3
         x(i)=y(i)
         y(i)=rdata(i)
      end do
      write(60) z
      goto 300
399   do i=1,256
         z(i)=0.3333*x(i)+0.6666*y(i)
      end do
      write(60) z
      rewind(60)
      close(40)
```

```
c*************************************************************
c                                                            *
c   High Frequency Enhancement                               *
c                                                            *
c*************************************************************
      print*,' HIGH FREQUENCY ENHANCEMENT '
      open(65,status='scratch',form='unformatted',
     +        recordtype='fixed',recl=1)
400   read(60,end=410) x
      en1=0.
      do i=1,256
         en1=en1+x(i)**2
      end do
      en1=sqrt(en1)
      write(65) en1
      goto 400
410   rewind(60)
      rewind(65)
      open(70,status='scratch',form='unformatted',
     +        recordtype='fixed',recl=256)
490   read(60,end=499) x
      read(65,end=499) en2
      if(en2.lt.5.e+5) then
         do i=1,80
            y(i)=x(i)
         end do
         do i=81,256
               y(i)=5.*x(i)
         end do
         else
         do i=1,23
            y(i)=x(i)
         end do
         do i=24,80
            y(i)=1.5*x(i)
         end do
         do i=81,256
            y(i)=7.*x(i)
         end do
      end if
      write(70) y
      j=j+1
      goto 490
499   rewind(65)
      rewind(70)
      close(60)
c*************************************************************
c                                                            *
c   Selection of Harmonics/Peaks of Voiced/Unvoiced Speech   *
c                                                            *
c*************************************************************
      print*,' SELECTION OF PEAKS/HARMONICS '
      open(80,status='scratch',form='unformatted',
```

```
     +         recordtype='fixed',recl=256)
500   read(70,end=599) x
      read(65,end=599) en2
      n=4
      if(en2.gt.5.e+5) then
         j=n+1
         do i=n+1,255
            amax=0.
            if(i.eq.j) then
               a=x(i-1)
               b=x(i)
               c=x(i+1)
               amax=max(a,b,c)
               if(amax.eq.b) y(i)=b
               if(amax.eq.a) then
               y(i-1)=a
                  j=i-1
                  else
               end if
               if(amax.eq.c) then
                  y(i+1)=c
                  j=i+1
                  else
               end if
               a=0.
               b=0.
               c=0.
               j=j+n
               else
            end if
         end do
c*************************************************************
c   Amplitude Threshold for Harmonics                         *
c*************************************************************
         a=60000.
         do i=1,256
            if(i.lt.21) then
               if(y(i).lt.a) y(i)=0.
               else
               b=a*(261-i)/240
               if(y(i).lt.b) y(i)=0.
            end if
         end do
         else
         call peak(x,y)
      end if
      write(80) y
      do i=1,256
         x(i)=0.
      end do
      goto 500
599   rewind(80)
      close(69)
      close(70)
```

```
c*************************************************************
c                                                            *
c    Synthesis of Speech using modified amplitudes,original*
c    Phase, and frequency.                                   *
c                                                            *
c*************************************************************
      print*,' SPEECH SYNTHESIS  '
      open(90,status='scratch',form='unformatted',
     +        recordtype='fixed',recl=256)
600   read(80,end=699)pk
      read(50,end=699)ph
      do t=1,256
         u(t)=0.
         do i=1,256
            amp=pk(i)
            phas=ph(i)
            if(amp.eq.0.) goto 610
            u(t)=u(t)+(amp*cos(((2*pi*(i-1)*(t-1))/512)+phas))
610         amp=0.
            phas=0.
         end do
      end do
      write(90) u
      goto 600
699   rewind(90)
      close(80)
      close(50)
c*************************************************************
c                                                            *
c   Averaging the data to original number of frames (N),and*
c   changing it to Integer*2 format.                         *
c                                                            *
c*************************************************************
      print*,' AVERAGING THE OUTPUT DATA TO N FRAMES '
      open(95,status='scratch',form='unformatted',
     +        recordtype='fixed',recl=256)
      j=1
700   read(90,end=799) rdata
      if(j.gt.1) goto 710
      do i=1,256
         x(i)=rdata(i)*(0.54-0.46*cos((2*pi/255)*(i-1)))
      end do
      j=j+1
      goto 700
710   k=mod(j,2)
      if(k.ne.0) goto 720
      do i=1,256
         y(i)=rdata(i)*(0.54-0.46*cos((2*pi/255)*(i-1)))
      end do
      do i=1,128
         z(i)=x(i)
         z(i+128)=(x(i+128)+y(i))
      end do
```

```
      write(95) z
      do i=1,256
         x(i)=0.
         u(i)=0.
      end do
      j=j+1
      goto 700
720   do i=1,256
        x(i)=rdata(i)*(0.54-0.46*cos((2*pi/255)*(i-1)))
      end do
      do i=1,128
         x(i)=x(i)+y(i+128)
      end do
      do i=1,256
         y(i)=0.
      end do
      j=j+1
      goto 700
799   do i=1,256
         z(i)=x(i)
      end do
      write(95) z
      rewind(95)
      close(90)
c*************************************************************
c                                                            *
c   Amplitude Noramlization                                  *
c                                                            *
c*************************************************************
      print*,' NORMALIZING THE OUTPUT WAVEFORM '
      open(110,status='scratch',form='unformatted',
     +         recordtype='fixed',recl=1)
      j=1
      a=0.
800   read(95,end=820) z
      amax=0.
      do i=1,256
         x(i)=abs(z(i))
         amax=max(amax,x(i))
      end do
      write(110) amax
      goto 800
820   j=j-1
      amax=0.
      rewind(110)
840   read(110,end=860) amax1
      amax=max(amax,amax1)
      goto 840
860   rewind(95)
      p=32760./amax
880   read(95,end=899) z
      do i=1,256
         s(i)=int(p*z(i))
      end do
```

```
      write(99) s
      goto 880
899   rewind(99)
      stop
      end
c************************************************************
c                                                           *
c   DFT Subroutine   (9:457)                                *
c                                                           *
c************************************************************
      subroutine fft(log2n,xr,xi,ntype)
      dimension xr(1),xi(1)
      integer  log2n,ntype,i,j,k,n,nv2,nm1,l,le,le1,ip
      real    xi,xr,tr,ti,ur,ui,wr,wi,ain,sign,pi
      data pi/3.14592653589793244/
      sign=-1.
      if(ntype.lt.0) sign=1.
      n=2**log2n
      nv2=n/2
      nm1=n-1
      j=1
      do 7 i=1,nm1
         if(i.ge.j) goto 5
         tr=xr(j)
         ti=xi(j)
         xr(j)=xr(i)
         xi(j)=xi(i)
         xr(i)=tr
         xi(i)=ti
5        k=nv2
6        if(k.ge.j) goto 7
         j=j-k
         k=k/2
         goto 6
7     j=j+k
      do 20 l=1,log2n
         le=2**l
         le1=le/2
         ur=1.
         ui=0.
         wr=cos(pi/le1)
         wi=sign*sin(pi/le1)
         do 20 j=1,le1
            do 10 i=j,n,le
               ip=i+le1
               tr=xr(ip)*ur-xi(ip)*ui
               ti=xr(ip)*ui+xi(ip)*ur
               xr(ip)=xr(i)-tr
               xi(ip)=xi(i)-ti
               xr(i)=xr(i)+tr
10             xi(i)=xi(i)+ti
            tr=ur*wr-ui*wi
            ti=ur*wi+ui*wr
```

```
             ur=tr
20           ui=ti
          if(ntype.gt.0) return
          ain=1./n
        do 30 i=1,n
           xr(i)=xr(i)*ain
30         xi(i)=xi(i)*ain
       return
       end
c***********************************************************
c                                                          *
c      subroutine PEAKS                                    *
c      Author: Flt.Lt. Nadeem A. Bashir                    *
c      Date  : February 1989                               *
c                                                          *
c***********************************************************
       subroutine peak(x,u)
       dimension x(256),u(*)
       integer i,j,k,npeak,n
       real x,u,a,b,c,amax
       npeak=0
       j=1
10     do 20 i=1,254
          a=x(i)
          b=x(i+1)
          c=x(i+2)
          if(a.lt.b.and.c.lt.b) then
             u(i)=b
             else
             u(i)=0.
             u(255)=0.
          end if
20     u(256)=0.
c**********************************************************
c                                                         *
c    Amplitude threshold for Peaks                        *
c                                                         *
c**********************************************************
       do i=1,256
          if(u(i).lt.4000.) u(i)=0.
       end do
       return
       end
```

Bibliography

1. Ahmed, M.S. "Comparison of Noisy Speech Enhancement Algorithms in Terms of LPC Perturbation," IEEE Transactions of Acoustics, Speech, and Signal Processing. ASSP-37 : 121-125 (January 1989).

2. Aschkenasy, Ernest and Mark R. Weiss MultiChannel Advanced Speech Enhancement development. Final Technical report No. RADC-TR-86-244. Rome Air Development Center, Griffis AFB, NY. December 1986.

3. Audio Data Conversion System. Maintenance Manual. D/N 00200-90151-A. Digital Sound Corporation, Santa Barbara, CA. December 1985.

4. Berouti, M., R. Schwartz, and J.Makhoul. " Enhancement of Speech Corrupted by Acoustic Noise, " IEEE Transactions of Acoustics, Speech, and Signal Processing. ASSP-27 : 208-211 (April 1979).

5. Boll, Steven F. " Suppression of Acoustic Noise in Speech using Spectral Subtraction, " IEEE Transactions of Acoustics, Speech, and Signal Processing. ASSP-27 : 113-120 (April 1979).

6. Bruselas, Capt Micheal A. Investigation of Speaker Independent Word Recognition using Multiple Features, Decision Mechanism, and Template Sets. MS Thesis, AFIT/GCE/ENG/86D-5. School of Engineering. Air Force Institute of Technology (AU), Wright-Patterson AFB OH. December 1986.

7. Harris, Fredric J. " On the use of Windows for Harmonic Analysis with the Discrete Fourier Transform, " Proceedings of the IEEE Vol.66 No.1 : 51-83 (January 1978)

8. Interactive Laboratory System. Introduction to ILS V6.0. Signal Technology Inc. Goleta, CA. 1986.

9. Kuc, Roman. " Introduction to Digital Signal Processing " New York : McGraw-Hill Inc. 1988.

10. Lim, Jae S. " Speech Enhancement ." New Jersey : Prentice-Hall Inc, 1983.

11. McAulay, R. J. and T. F. Quatieri. " Magnitude-only Reconstruction using a Sinusoidal Speech Model, " IEEE Transactions of Acoustics, Speech, and Signal processing. : 27.6.1-27.6.4 (March 1984).

12. McAulay, Robert J. and Thomas F. Quatieri. " Speech Transformation Based on a Sinusoidal Representation" IEEE Transaction of Acoustics, Speech, and Signal Processing. : 489-492 (March 1985).

13. Kabrisky, Mathew, Professor. Personnel Discussion. School of Engineering, AFIT (AU), Wright-Patterson AFB OH. January 1989.

14. Parsons, Thomas W. " Separation of Speech from Interfering Speech by means of Harmonic Selection," The Journal of the Acoustical Society of America , Volume 60 : 911-918 (October 1976).

Vita

Nadeem A. Bashir [redacted] He graduated from Aitchison College Lahore, Pakistan, in 1975. In January 1980 he graduated, with distinction, from the Pakistan Air Force College of Aeronautical Engineering with the degree of Bachelor of Avionics Engineering. He entered the School of Engineering, Air Force Institute of Technology in June 1987.

[redacted]

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

Form Approved OMB No. 0704-0188

1a. REPORT SECURITY CLASSIFICATION UNCLASSIFIED	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION/AVAILABILITY OF REPORT APPROVED FOR PUBLIC RELEASE DISTRIBUTION UNLIMITED
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUM' ER(S) AFIT/GE/ENG/89M-2	5. MONITORING ORGANIZATION REPORT NUMBER(S)
6a. NAME OF PERFORMING ORGANIZATION School of Engineering; 6b. OFFICE SYMBOL (If applicable) AFIT/ENG	7a. NAME OF MONITORING ORGANIZATION
6c. ADDRESS (City, State, and ZIP Code) Air Force Institute of Technology Wright-Patterson AFB, OH 45433	7b. ADDRESS (City, State, and ZIP Code)
8a. NAME OF FUNDING/SPONSORING ORGANIZATION RADC/IRAA Griffis AFB; 8b. OFFICE SYMBOL (If applicable)	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
8c. ADDRESS (City, State, and ZIP Code) Capt. David B.Stockton RADC/IRAA Griffis AFB, NY 13441-5700 AV 587-4024; AC 315-330-4024	10. SOURCE OF FUNDING NUMBERS: PROGRAM ELEMENT NO. / PROJECT NO. / TASK NO. / WORK UNIT ACCESSION NO.

11. TITLE (Include Security Classification)
Phoneme Adjustment in Enhanced Speech UNCLASSIFIED

12. PERSONAL AUTHOR(S)
Bashir, Nadeem A. Flt.Lt. Pakistan Air Force

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Year, Month, Day)	15. PAGE COUNT
MS Thesis	FROM ______ TO ______	March 1989	68

16. SUPPLEMENTARY NOTATION

17. COSATI CODES			18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
FIELD	GROUP	SUB-GROUP	Speech Processing, Amplitude Spectrum, Spectrum Smoothing
17	02		Harmonics/Peaks Selection, Speech Reconstruction

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

Thesis Chairman : Mathew Kabrisky, PhD
Professor of Electrical Engineering

Cont/-

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT ☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT. ☐ DTIC USERS	21. ABSTRACT SECURITY CLASSIFICATION UNCLASSIFIED
22a. NAME OF RESPONSIBLE INDIVIDUAL Dr. Mathew Kabrisky Professor, GS-15	22b. TELEPHONE (Include Area Code) (513) 255-5276; 22c. OFFICE SYMBOL AFIT/ENG

DD Form 1473, JUN 86 Previous editions are obsolete. SECURITY CLASSIFICATION OF THIS PAGE

UNCLASSIFIED

continued from block 19: Abstract

A system has been developed to enhance the quality and intelligibility of speech which had been pre-processed by a Speech Enhancement Unit (SEU) at Rome Air Development Center, Griffis AFB. The system processes the speech in the frequency domain using 512-point DFT. The amplitude spectrum of voiced regions of speech is smoothed in order to reduce the effects of noise. Frequencies above 2.5 KHz are enhanced as they had been attenuated by SEU. Harmonics of the glottal pitch frequency of voiced speech and peaks of amplitude spectrum for unvoiced speech are selected to further reduce eliminate the noisy components from the spectrum. The harmonics selected are not necessarily the exact harmonics of the glottal frequency. The two neighboring frequency points of the harmonics are checked and the maximum of those three points are selected. All the harmonics/peaks selected are compared against a threshold and values below the threshold are deleted. Speech is then reconstructed using amplitude, phase, and frequency of the harmonics/peaks selected. The number of harmonics/peaks used for reconstruction varied from 40 to 50 in each frame of 256 DFT points. The reconstructed speech has much better quality and improved SNR.